

6th
International
Global Wordnet
Conference

GWC2012

Proceedings

GWC2012
Global Wordnet Conference
January 9-13, 2012 • Matsue, Japan

GWC 2012

6th International Global Wordnet Conference



This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks.
Duplication of this publication or parts thereof is permitted only under the provisions of the Czech Copyright Law, in its current version, and permission for use must always be obtained from the Global Wordnet Association . Violations are liable for prosecution under the Czech Copyright Law.

Editors © Christiane Fellbaum, Piek Vossen, 2012
This edition © Tribun EU, Brno, 2012

ISBN 978-80-263-0244-5

Program Committee List

Eneko Agirre	University of the Basque Country, San Sebastian, Spain
Pushpak Bhattacharyya	IIT Bombay, India
Sonja Bosch	Department of African Languages, University of South Africa, Pretoria, South-Africa
Agata Cybulska	Department of Linguistics and Communication/VU University Amsterdam, Amsterdam, Netherlands
Christiane Fellbaum	Princeton University, Princeton, US
Darja Fišer	University of Ljubljana, Ljubljana, Slovenia
Yoshihiko Hayashi	Osaka University, Osaka, Japan
Axel Herold	Berlin-Brandenburg Academy of Sciences, Berlin, Germany
Ales Horak	Masaryk University, Faculty of Informatics, Brno, Czech Republic
Chu-Ren Huang	Hong Kong Polytechnic University, Hong Kong, China
Hitoshi Ishara	Toyohashi University of Technology, Toyohashi, Japan
Toru Ishida	Department of Social Informatics, Kyoto University, Kyoto, Japan
Kyoko Kanzaki	National Institute of Japanese Language and Linguistics, Kyoto, Japan
Adam Kilgarriff	Lexical Computing Ltd, UK.
Kow Kuroda	Waseda University, Tokyo, Japan
Andrea Marchetti	Istituto di Informatica e Telematica – CNR, Pisa, Italy
Monica Monachini	Istituto di Linguistica Computazionale – CNR, Pisa, Italy
Naoyuki Ono	Tohoku University, Tohoku, Japan
Heili Orav	University of Tartu, Tartu, Estonia
Karel Pala	Masaryk University, Brno, Czech Republic
Adam Pease	Articulate Software, California, US
Ted Pedersen	University of Minnesota, Duluth, US
Maciej Piasecki	Institute of Applied Informatics, Wrocław University of Technology, Warchau, Poland
German Rigau	University of the Basque Country, San Sebastian, Spain
Horacio Rodriguez	Universitat Politecnica de Catalunya, Barcelona, Spain
Peter Schulam	Carnegie-Mellon University, US
Claudia Soria	Istituto di Linguistica Computazionale – CNR, Pisa, Italy
Virach Sornlertlamvanich	National Electronics and Computer Technology Center, Bangkok, Thailand
Takenobu Tokunaga	Tokyo Institute of Technology, Tokyo, Japan
Zygmunt Vetulani	Adam Mickiewicz University, Poznan, Poland
Piek Vossen	Vrije Universiteit Amsterdam, Amsterdam, Netherlands
Daniela Katunar	University of Zagreb, Croatia

List of Additional Reviewers

A

Agerri, Rodrigo

F

Frontini, Francesca

O

Oliver, Antoni

V

Vare, Kadri

Message from conference chairs

The Sixth Global WordNet Conference testifies to the continued lively interest in the creation of new wordnets, wordnet tools and applications. We are excited to welcome several new languages into the global wordnet community, to learn about novel methods for wordnet construction and enhancements and to discuss challenges for representing words and meanings posed by languages whose lexicons had not previously been examined and represented as a semantic network.

Japan is a country well known for its natural beauty, gracious hospitality and organizational efficiency. All three conspired to ensure that ours will be a successful meeting. We are grateful to the local organizers for their hard work and the Shimane Prefecture and Matsue City for their sponsorship. The Program Committee members performed their voluntary duties responsibly and timely. Finally, we are honored to have two distinguished Japanese colleagues as our guest speakers, whose presence and presentations will inspire us.

Christiane Fellbaum and Piek Vossen

Table of Contents

Leveraging Sentiment to Compute Word Similarity	
<i>Balamurali A R, Subhabrata Mukherjee, Akshat Malu and Pushpak Bhattacharyya.....</i>	10
Toward Building a Large-Scale Arabic Sentiment Lexicon	
<i>Muhammad Abdul-Mageed and Mona Diab.....</i>	18
WordNet Atlas: a web application for visualizing WordNet as a zoomable map	
<i>Matteo Abrate, Clara Bacciu, Andrea Marchetti and Maurizio Tesconi.....</i>	23
Verbs in Sanskrit Wordnet	
<i>Tanuja Ajotikar, Malhar Kulkarni and Pushpak Bhattacharyya.....</i>	30
Using WordNet to Handle the OOV Problem in English to Bangla Machine Translation	
<i>Khan Md. Anwarus Salam, Setsuo Yamada and Tetsuro Nishino.....</i>	35
A new hierarchy of ArchiWordNet (AWN): building parts implementation with image	
<i>Anna Rita Bertorello.....</i>	40
Introduction to Gujarati WordNet	
<i>Brijesh Bhatt, C Bhensdadia, Pushpak Bhattacharyya, Dinesh Chauhan and Kirit Patel.....</i>	45
Extending CzechWordNet Using a Bilingual Dictionary	
<i>Marek Blahuš and Karel Pala.....</i>	50
Japanese SemCor: A Sense-tagged Corpus of Japanese	
<i>Francis Bond, Timothy Baldwin, Richard Fothergill and Kiyotaka Uchimoto.....</i>	56
A Survey of WordNets and their Licenses	
<i>Francis Bond and Kyonghee Paik.....</i>	64
Semantic Hand-Tagging of the SenSem Corpus Using Spanish WordNet Senses	
<i>Irene Castellón, Salvador Climent, Marta Coll-Florit, Marina Lloberes and German Rigau</i>	72
A Study of the Sense Annotation Process: Man v/s Machine	
<i>Arindam Chatterjee, Salil Joshi, Pushpak Bhattacharyya, Diptesh Kanodia and Akhlesh Meena</i>	79
A Computer Aided Approach for Enriching WordNet with Semantic Definition	
<i>Lin Dai, Weitao Zhou and Heyan Huang</i>	86
Combining Wordnet and Crosslingual multi-terminology health portal to access health information	
<i>Stefan J Darmoni, Julien Grosjean, Tayeb Merabti, Nicolas Griffon, Badisse Dahamna and Dominique Dutoit.....</i>	94
Revisiting a Brazilian WordNet	
<i>Valeria De Paiva and Alexandre Rademaker.....</i>	100
Scalar Properties of Emotion Verbs and Their Representation in WordNet	
<i>Christiane Fellbaum and Yvette Yannick Mathieu</i>	105
sloWNet 3.0: development, extension and cleaning	
<i>Darja Fišer, Jernej Novak and Tomaž Erjavec</i>	113
Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base	
<i>Aitor Gonzalez, Egoitz Laparra and German Rigau.....</i>	118

New WordNet-based semantic relatedness measurement: Using new information content metric and k-means clustering algorithm	126
<i>Mohamed Ali Hadj Taieb and Mohamed Ben Aouicha</i>	
Computing Cross-Lingual Synonym Set Similarity by Using Princeton Annotated Gloss Corpus	
<i>Yoshihiko Hayashi</i>	134
Restructuring Adjectives in WordNet with ClusterEditor	
<i>Isaac Julien and Christiane Fellbaum</i>	142
An Extractive Approach of Text Summarization of Assamese using WordNet	
<i>Chandan Kalita, Navanath Saharia and Utpal Sharma</i>	149
ASSAMESE VOCABULARY AND ASSAMESE WORDNET BUILDING: AN ANALYSIS	
<i>Shikhar Kr. Sarma, Utpal Saikia, Mayashree Mahanta and Himadri Bharali</i>	155
Encoding Commonsense Lexical Knowledge into WordNet	
<i>Gianluca Lebani and Emanuele Pianta</i>	159
Visual Study of Estonian Wordnet using Bipartite Graphs and Minimal Crossing algorithm	
<i>Ahti Lohk, Kadri Vare and Leo Vohandu</i>	167
Rethinking WordNet's Domains	
<i>Xiaojuan Ma and Christiane Fellbaum</i>	173
An Implementation of a System of Verb Relations in plWordNet 2.0	
<i>Marek Maziarz, Maciej Piasecki and Stanislaw Szpakowicz</i>	181
Approaching plWordNet 2.0	
<i>Marek Maziarz, Maciej Piasecki and Stanislaw Szpakowicz</i>	189
NOUNS IN ODIA: AN ONTOLOGICAL PERSPECTIVE	
<i>Panchanan Mohanty</i>	197
Ontology of Sanskrit Wordnet: Nouns and Verbs	
<i>Sanghamitra Mohanty and Das Adhikary</i>	205
Using WordNet to predict numeral classifiers in Chinese and Japanese	
<i>Hazel Mok, Eshley Gao and Francis Bond</i>	211
Mapping a corpus-induced ontology of action verbs on ItalWordNet	
<i>Massimo Moneglia, Alessandro Panunzi, Gloria Gagliardi, Monica Monachini, Irene Russo and Francesca Frontini</i>	219
Using a Bilingual Resource to Add Synonyms to a Wordnet: FinnWordNet and Wikipedia as an Example	
<i>Jyrki Niemi, Krister Lindén and Mirka Hyvärinen</i>	227
Building WordNets by machine translation of sense tagged corpora	
<i>Antoni Oliver and Salvador Climent</i>	232
Kannada Verbs and their Automatic Sense Disambiguation	
<i>S Parameswarappa and V.N Narayana</i>	240
Wordnet and SUMO for Sentiment Analysis	
<i>Adam Pease, John Li and Karen Nomorosa</i>	248

Linking and Validating Nordic and Baltic Wordnets - A Multilingual Action in META-NORD <i>Bolette Sandford Pedersen, Lars Borin, Markus Forsberg, Krister Lindén, Heili Orav and Eirikur Rögnvaldsson</i>	254
Using WordNet into UKB in a Question Answering System for Basque <i>Olatz Perez De Viñaspre Garralda, Maite Oronoz Anchordoqui and Olatz Ansa Osteriz</i>	260
Automated Mapping of Polish Proper Names on plWordNet Hypernymy Structure <i>Maciej Piasecki, Roman Kurc and Agnieszka Indyka-Piasecka</i>	268
Automated Generation of Derivative Relations in the Wordnet Expansion Perspective <i>Maciej Piasecki, Radosław Ramocki and Marek Maziarz</i>	273
Mass noun classifiers in Nepali <i>Madhav Pokharel</i>	281
Finding a Location for a New Word in WordNet <i>Paula Pääkkö and Krister Lindén</i>	286
Introducing WordNet in Interpreting Studies - Implications and Desiderata <i>Francesca Quattri and Chu-Ren Huang</i>	294
Low-cost ontology development <i>Adam Rambousek and Marek Grác</i>	299
Migrating Cornetto Lexicon to New XML Database Engine <i>Adam Rambousek and Aleš Horák</i>	305
A Proposed Nepali Synset Entry and Extraction Tool <i>Arindam Roy, Sunita Sarkar and Bipul Shyam Purkayastha</i>	312
Automatic Extension of WOLF <i>Benoit Sagot and Darja Fišer</i>	317
A Novel Approach for Document Classification using Assamese WordNet <i>Jumi Sarmah, Navanath Saharia and Shikhar K Sarma</i>	324
Refining WordNet adjective dumbbells using intensity relations <i>Vera Sheinman, Takenobu Tokunaga, Isaac Julien, Peter Schulam and Christiane Fellbaum</i>	330
The C-Cat Wordnet Package: An Open Source Package for modifying and applying Wordnet <i>Keith Stevens, Terry Huang and David Buttler</i>	338
Linking WordNet to DBpedia <i>Aynaz Taheri and Mehrnoush Shamsfard</i>	344
Extension of Phrases for Article Determination using WordNet Thesaurus <i>Hiromi Takeuchi, Hirofumi Miyake, Atsuo Kawai, Ryo Nagata and Hokuto Ototake</i>	349
Linking specific and generalist knowledge – Building terminological resources from catalogues and generalist resources – <i>Benoît Trouvilliez</i>	357
A Linguistic Approach to Feature Extraction Based on a Lexical Database of the Properties of Adjectives and Adverbs <i>Franco Tuveri and Manuela Angioni</i>	365

Teaching WordNet to Sing like an Angel and Cry like a Baby: Learning Affective Stereotypical Behaviors from the Web <i>Tony Veale</i>	371
Multiword Verbs in WordNets <i>Veronika Vincze, Attila Almási and János Csirik</i>	377
Cross-lingual event-mining using wordnet as a shared knowledge interface <i>Piek Vossen, Aitor Soroa, Beñat Zapirain and German Rigau</i>	382
Upgrading WordNet: a Terminological Point of View <i>Cristian Zanotti, Jorge Vivaldi and Mercè Lorente Casafont</i>	390
SENEQA – System for Quality Testing of Wordnet Data <i>Tomáš Čapek</i>	400
Author Index	405
Keyword Index.....	413

Invited Talk

Eventiveness and Argument Selection in Nominals

Naoyuki Ono (Tohoku University)

Verbs are a prototypical argument-taking category. They need to select proper arguments because the event denoted by a verb is represented in terms of semantic roles that the arguments bear in relation with each other. Thus, the argument selection of verbs is in a substantial sense a linguistic manifestation of eventiveness. Is this the case in other argument-taking categories? Obviously nouns are an argument-taking category, but the expression of eventiveness is not as self-evident as with verbs because they denote individuals as well as events. Consequently, do they select arguments in a manner substantially different from verbs? If so, how is eventiveness concerned with the syntactic or lexical realization of semantic relations associated with the arguments of nouns?

In this talk I argue that there are two distinct modes of argument selection in nominals depending on how the denotation of the head noun is determined with reference to ontological relations encoded in lexical structure. Along with Pustejovsky's basic idea that the distinction between individual-level predicates (e.g. *Firemen are intelligent.*) and stage-level predicates (e.g. *Firemen are available.*) can be extended to the distinction in nominals, I assume that nominals generally fall into two semantic categories in terms of eventiveness: individual-level and stage-level nominals.

Japanese has a wide variety of noun-forming affixes that refer to this distinction. For instance, complex agentive nouns with affixes such as *-sya* (-者) and *-te* (-手) are stage-level (e.g. *hokoosya* 'pedestrian' and *utaite* 'singer') whereas those affixed with *-syu* (-手) and *-ka* (-家) (e.g. *untensyu* 'driver' and *syoosetuka* 'novelist') are individual-level. Object-denoting complex nouns with the affixes *-mono*, *-butu* (-物) fall into the two classes as well (e.g. *wasuremono* 'lost article' vs. *nisemono* 'imitation'). With an extensive survey of noun-forming affixes in Japanese, I will discuss how the semantic distinction in question is implemented in argument selection of nominals. Specifically I propose that there are two different modes of argument selection in nominals, which I will call the eventive mode and the relational mode, and selection of arguments by stage-level nominals is conducted via the eventive mode whereas selection of arguments by individual-level nominals is done in the relational mode.

Leveraging Sentiment to Compute Word Similarity

Balamurali A R^{†,‡} Subhabrata Mukherjee[†] Akshat Malu[‡] Pushpak Bhattacharyya[‡]

[†] IITB-Monash Research Academy, IIT Bombay

[‡]Dept. of Computer Science and Engineering, IIT Bombay

{balamurali,subhabratam,akshatmalu,pb}@cse.iitb.ac.in

Abstract

In this paper, we introduce a new WordNet based similarity metric, *SenSim*, which incorporates sentiment content (*i.e.*, degree of positive or negative sentiment) of the words being compared to measure the similarity between them. The proposed metric is based on the hypothesis that knowing the sentiment is beneficial in measuring the similarity. To verify this hypothesis, we measure and compare the annotator agreement for 2 annotation strategies: 1) sentiment information of a pair of words is considered while annotating and 2) sentiment information of a pair of words is not considered while annotating. Inter-annotator correlation scores show that the agreement is better when the two annotators consider sentiment information while assigning a similarity score to a pair of words.

We use this hypothesis to measure the similarity between a pair of words. Specifically, we represent each word as a vector containing sentiment scores of all the content words in the WordNet gloss of the sense of that word. These sentiment scores are derived from a sentiment lexicon. We then measure the cosine similarity between the two vectors. We perform both intrinsic and extrinsic evaluation of *SenSim* and compare the performance with other widely used WordNet similarity metrics.

1 Introduction

Use of similarity metrics is unavoidable in many Natural Language Processing (NLP) systems. They form core of many NLP tasks like Word Sense disambiguation (Banerjee & Pedersen 2002), malapropism detection (Hirst & St-Onge

1997), context sensitive spelling correction (Patwardhan 2003) *etc.* The underlying principle of these metrics has been distributional similarity in terms of their meaning. For example, *refuge* and *asylum* are similar words because they have the same meaning and similar set of words accompany them in a given context. Based on the meaning alone, these words are mutually replaceable.

At present, there are various advanced text editors which have the ability to replace a word based on the meaning suitable for the domain/genre in which the article is being written. To select an appropriate replacement word, they follow a similarity based on meaning alone. Motivated by the idea of sub-languages (Grishman 2001), we believe similarity based on meaning alone cannot suffice this need. For example, in the previous case, even though *refuge* can be replaced with *asylum*, *mad house* cannot be used to do so. This is because *mad house* evokes a negative connotation or sentiment which makes the word unsuitable for replacement.

In this paper, we propose a new WordNet based similarity metric, *SenSim*, which takes into consideration the sentiment content (*i.e.*, degree of positive or negative sentiment) of the words being compared. We create vector representations of WordNet glosses and compare their cosines to calculate the similarity score. To include the sentiment content of the words being compared, we include sentiment scores of the content words of the gloss into the vector. The main contribution of this paper is in addressing the following question:

Does inclusion of sentiment content as an additional parameter for comparison improve similarity measurement?

We perform an intrinsic and extrinsic evaluation of the *SenSim* metric. As a part of intrinsic evaluation, we manually annotate a set of 48 random word pairs based on their word similarity on a scale of 1-5 and calculate the correlation

between our metric and the annotator scores. We also calculate the correlation between three popular WordNet based similarity metrics and the annotated dataset (gold standard). Our results show that the new metric has a better correlation with the annotator scores than the other metrics used for this study.

For extrinsic evaluation, we compare the effect of SenSim metric *to mitigate the effect of unknown feature problem in supervised sentiment classification using synset replacement strategy* as proposed by Balamurali et al. (2011). A classifier performs with lesser accuracy on a test set than on a training set due to the presence of features not seen during its training phase. To alleviate this, they propose to replace features not present in test set with *similar* features present in the training set using a similarity metric thereby reducing the difference between the feature distribution of the training and the test set. Our results show that SenSim based document level sentiment classifier performs better than other WordNet based metrics.

We should mention that this work is also motivated by the concept of semantic prosody (Sinclair 2004). According to corpus linguistics, semantic prosody is described as a property of a word and a feature that distinguishes near-synonyms (Partington 2004). It explicitly assigns a word with its positive and negative attitudinal meanings. It is also a gradable property with a word having a gradable 'favourable' or 'unfavourable' prosody depending on how frequently it occurs in good, bad or neutral context. In this context, the development of Sen-Sim is an attempt to integrate information about semantic prosody into the WordNet ontology.

The rest of the paper is organized as follows: Section 2 describes the related work and distinguishes our work from the existing similarity metrics. We describe our metric and related terminologies in Section 3. We explain evaluation strategy in Section 4. Section 5 describes existing similarity metrics that we use for comparison. Experimental setup is given in Section 6. Results of the experiments are discussed and analyzed in Section 7. Section 8 concludes the paper.

2 Related work

Various approaches for evaluating the similarity between two words can be broadly classified into two categories: edge-based methods and information-content-based methods. One of the

earliest works in edge-based calculation of similarity is by Rada et al. (1989), where in, they propose a metric "Distance" over a semantic net of hierarchical relations as the shortest path length between the two nodes. This has been the basis for all the metrics involving simple edge-counting to calculate the distance between two nodes. However, the simple edge-counting fails to consider the variable density of nodes across the taxonomy. It also fails to include relationships other than the 'is-a' relationship, thus, missing out on important information in a generic semantic ontology, like WordNet.

In contrast to edge-based methods, Richardson et al. (1994) and Resnik (1995a) propose a node-based approach to find the semantic similarity. They approximate conceptual similarity between two WordNet concepts as the maximum information content among classes that subsume both the concepts. Resnik (1995b) advanced this idea by defining the information content of a concept based on the probability of encountering an instance of that concept. Alternatively, Wu & Palmer (1994) compare two concepts based on the length of the path between the root of the hierarchy and the least common subsumer of the concepts.

Jiang & Conrath (1997) and Leacock et al. (1998) combine the above two approaches by using the information content as weights for the edges between concepts. They further reinforce the definition of information content of a concept by adding corpus statistical information.

Instead of measuring the similarity of concepts, some other approaches measure their relatedness. Hirst & St-Onge (1997) introduce an additional notion of direction along with the length of paths for measuring the relatedness of two concepts. Banerjee & Pedersen (2003) and Patwardhan (2003) leverage the gloss information present in WordNet in order to calculate the relatedness of two concepts. Banerjee & Pedersen (2003) assigns relatedness scores based on the overlap between the gloss of the two concepts. Patwardhan (2003) use a vector representation of the gloss, based on the context vector of the terms in the gloss. The relatedness is then the cosine between the gloss vectors of the two concepts.

Our work is most related to the work of Wan & Angryk (2007) which improves on Banerjee & Pedersen (2003) and Patwardhan (2003) by including relations other than the is-a relationship.

They use an extended gloss definition for a concept which is defined as the original gloss appended by the gloss of all the concepts related to the given concept. They create concept vectors for each sense based on which they create context vectors which are an order higher to the concept vectors. Finally, they use cosine of the angle between the vectors of the different concepts to find their relatedness. This approach is better than other approaches as it captures the context of the concepts to a much larger extent. However, all these methods lack on a common ground. They fail to incorporate sentiment information in calculating the similarity/relatedness of two concepts. We postulate that sentiment information is crucial in finding the similarity between two concepts.

3 SenSim metric

The underlying hypothesis that we follow for creating this metric is that *knowing the sentiment is beneficial in measuring the similarity*. In order to implement a metric based on this hypothesis, we incorporate sentiment values of the words being compared. Similar to Wan & Angryk (2007), we follow WordNet gloss based comparison technique to develop this metric. Gloss based technique has an inherent advantage over edge-based and information-content-based metrics as it is applicable to all POS categories without any distinction.

3.1 Gloss vector

We represent gloss of a synset in the form of a vector, we define this vector as gloss vector. To obtain the gloss of the words being compared, the corresponding sense used for each word needs to be known. We assume that we are provided with the synset corresponding to each word that needs to be compared. In other scenarios where the synset corresponding to word is not given, a close approximation can be taken by using its respective WordNet first sense. Each dimension of the gloss vector represents a sentiment score of the respective content word. To obtain the sentiment scores, we use an external sentiment lexicon and assign sentiment values based on different scoring functions.

We use SentiWordNet 1.0¹ as the external sentiment lexicon to incorporate the sentiment values in the gloss vector. This WordNet based resource has polarity scores attached to synsets (Esuli &

Sebastiani 2006). Each synset in this resource is marked with 3 scores: a positive score, a negative score and an objective score, with the scores summing up to 1. As the sentiment scores are attached to synsets rather than lexemes, we disambiguate the WordNet gloss to obtain the corresponding synsets. Based on synsets thus found, we assign sentiment scores to each dimension of the gloss vector.

Representing gloss in the form of vectors is not new, but novelty of our approach is in the incorporation of sentiment score to each dimension of the gloss vector.

3.2 Augmenting gloss vector

Gloss contains a few content words averaging between 5-7. This creates a sparse vector space. To reduce the sparseness of the gloss vector, we augment the original gloss with the gloss of the *related synsets*. We use different WordNet relations , based on the POS category, to find the related synsets. Apart from the *adjacent* related synset, we add more context by further expanding the related synsets using synsets of content words of the original gloss.

Not all WordNet relations can be used for the expansion procedure as degree of sentiment content may change or not get carried to the next level. By taking relative transfer of the sentiment content from one synset to another for different WordNet relations, we empirically found a set of WordNet relations suitable for each POS category. Details of the WordNet relations used for expansion process are given in Table 1.

POS	WordNet relations used for expansion
Nouns	<i>hypernym, hyponym, nominalization</i>
Verbs	<i>nominalization, hypernym, hyponym</i>
Adjectives	<i>also see, nominalization, attribute</i>
Adverbs	<i>derived</i>

Table 1: WordNet relations used for enhancing the context of Gloss vector

3.3 Scoring function

As SentiWordNet provides 3 scores for each synset: positive score, negative score and objective score, we devise a scoring function to capture the sentiment content of the words as a single real value. We explain four variants of the scoring function used in this study below:

¹<http://sentiWordNet.isti.cnr.it/>

Sentiment Difference (SD)

Difference between the positive score and the negative score is taken as the sentiment score of the synset concerned.

$$Score_{SD}(A) = SWN_{pos}(A) - SWN_{neg}(A)$$

Here $SWN_{pos}(A)$ signifies the positive score pertaining to the synset A . The sign of the score represents the orientation of the sentiment. If the sign is negative, the word has a negative connotation otherwise it has a positive connotation. If the value is zero, it is objective in nature.

Sentiment Max (SM)

For this function, we use the greater of the positive or negative score of the synset as sentiment score of the synset concerned.

$$Score_{SM}(A) = \max(SWN_{pos}(A), SWN_{neg}(A))$$

The orientation of the word is again distinguished by the sign. For negative connotation, a negative of the score returned is used else a positive score is taken.

Sentiment Threshold Difference (TD)

As the gloss is represented in the form of a vector with each dimension representing a sentiment score, dimensions which have zero magnitude may not contribute to the sentiment content but the presence of each dimension is necessary for overall similarity computation. In order to avoid such scenarios, we introduce a threshold value that ensures, in case of objective words, a zero value is never encountered. We take a threshold value of 1 to compute the variant of SD scoring function.

$$Score_{TD}(A) = \frac{\text{sign}(SWN_{pos}(A) - SWN_{neg}(A)) * (1 + \text{abs}(SWN_{pos}(A) - SWN_{neg}(A)))}{(1 + \text{abs}(SWN_{pos}(A) - SWN_{neg}(A)))}$$

Sentiment Threshold Max (TM)

SM scoring function is modified to handle zero magnitude problem as explained above.

$$Score_{TM}(A) = \frac{\text{sign}(\max(SWN_{pos}(A), SWN_{neg}(A))) * (1 + \text{abs}(\max(SWN_{pos}(A), SWN_{neg}(A))))}{(1 + \text{abs}(\max(SWN_{pos}(A), SWN_{neg}(A))))}$$

3.4 Computing similarity

To compute the similarity between two word pairs, we take the cosine similarity of their corresponding gloss vectors.

$$\begin{aligned} SenSim_x(A,B) &= \cosine(gloss_{vec}(\text{sense}(A)), \\ &\quad gloss_{vec}(\text{sense}(B))) \\ \text{where} \\ gloss_{vec} &= 1:score_x(1) \quad 2:score_x(2) \\ &\dots \dots n:score_x(n) \\ score_x(Y) &= \text{Sentiment score of word } Y \\ x &= \text{Scoring function of type} \\ &SD/SM/TD/TM \end{aligned}$$

4 Evaluation

To evaluate SenSim, we follow two methodologies: an intrinsic evaluation and an extrinsic evaluation. For intrinsic evaluation, we compare the correlation of the metric with a gold standard dataset. We also compare correlation of different existing similarity metrics with this gold standard and show how our new metric performs. To perform an extrinsic evaluation, we use SenSim metric to mitigate the effect of the unknown feature problem in supervised sentiment classification. Details of the same are explained in the following subsections.

4.1 Intrinsic evaluation: correlation with human annotators

We develop a new similarity dataset and manually annotate them with similarity scores. We did not use existing similarity datasets as we require the correct sense of the words being compared.

Dataset for annotation task

We chose 48 random word pairs for this task with each POS category having 12 word pairs each. Sentences, containing these words, are constructed to get the exact sense of these word pairs. Based on these sentences, words are sense disambiguated using WordNet 2.1 to get the corresponding synsets. A part of the word pairs used in this paper are given in Table2 ². Each word pair has a WordNet synset offset, prefixed with the POS, category attached to it.

Word Pairs	
regular_42374008	accustomed_426235
tardily_3100974	lately_3108293
randomly_371128	specifically_341621
pretentious_41915502	arrogant_41957189
defense_1811665	attack_1958708

Table 2: Sample word pairs from dataset used for experiments

²The complete set of word pairs is not included due to lack of space but is available on request

Annotation strategy

To test our hypothesis that *incorporation of sentiment into a similarity metric gives a better result than comparing similarity based on meaning alone*, we perform a similarity annotation task between two human annotators. The annotators were asked to annotate word pairs using two different strategies.

1. *Annotation based on Meaning* : Instructions were given to each annotator to give a score between 1-5 to word pairs based on semantic similarity with a score of 5 representing synonymous word pairs and 1 representing no relation between the words.
2. *Annotation based on Sentiment and Meaning combined* : In this case, the annotators were asked to rate on a scale between 1-5 whether words were interchangeable given a similar context and similar sentiment content. A score of 5 implies perfect interchangeability.

The dataset generated thus forms our gold standard data. Correlation between the annotator scores is taken to test the hypothesis. As a part of this evaluation, we also take the correlation scores between different existing similarity metrics with the gold standard data, and compare the respective correlation scores with SenSim.

4.2 Extrinsic evaluation: synset replacement using similarity metrics

In this section, we evaluate SenSim metric using an application of similarity metrics, and compare the application performance with that of widely used similarity metrics. We use the synset replacement strategy using similarity metrics by Balamurali et al. (2011) for evaluating our metric. In this study, the authors showed that similarity metrics can be used to mitigate the effect of unseen feature problem in supervised classification. The objective of their study is to classify documents based on their sentiment content into positive class or negative class. Each document to be classified is represented as a group of synsets obtained after sense disambiguation of the content words of the document. A document level sentiment classifier is created based on the training corpus and its performance is measured on the test corpus. A trained classifier will not perform with similar/same accuracy if new features are found in the test corpus.

```

Input: Training Corpus, Test Corpus,  

        Similarity Metric  

Output: New Test Corpus  

T:= Training Corpus;  

X:= Test Corpus;  

S:= Similarity metric;  

train.concept.list = get_list.concept(T) ;  

test.concept.list = get_list.concept(X) ;  

for each concept C in test.concept.list do  

    temp_max_similarity = 0 ;  

    temp.concept = C ;  

    for each concept D in train.concept.list do  

        similarity_value = get_similarity_value(C,D,S);  

        if (similarity_value > temp_max_similarity)  

            then  

                temp_max_similarity= similarity_value;  

                temp.concept = D ;  

        end  

    end  

    C = temp.concept ;  

    replace_synset_corpus(C,X);  

end  

Return X ;

```

Algorithm 1: Synset replacement using similarity metric

To mitigate this effect, they follow a *replacement strategy*. In this strategy, if a synset encountered in a test document is not found in the training corpus, it is replaced by one of the synsets present in the training corpus. The substitute synset is determined on the basis of its similarity with the concerned synset in the test document. The synset that is replaced is referred to as an *unseen synset* as it is not known to the trained model.

Algorithm 1 shows the replacement algorithm devised by Balamurali et al. (2011). The algorithm follows from the fact that the similarity value for a synset with itself is maximum. Similarity metrics used in this study are explained in the following section.

5 Metrics used for comparison

We compare SenSim with three existing similarity metric. They are:

LIN: The metric by Lin (1998) uses the information content individually possessed by two concepts in addition to that shared by them. The information content shared by two concepts A and B is given by their most specific subsumer (lowest super-ordinate(*lso*)). Thus, this metric defines the similarity between two concepts as

$$sim_{LIN}(A, B) = \frac{2 \times \log Pr(lso(A, B))}{\log Pr(A) + \log Pr(B)}$$

Lesk: Each concept in WordNet is defined

Annotation Strategy	Overall	NOUN	VERB	ADJECTIVES	ADVERBS
Meaning	0.768	0.803	0.750	0.527	0.759
Meaning + Sentiment	0.799	0.750	0.889	0.720	0.844

Table 3: Pearson correlation coefficient between two annotators for various annotation strategy

through gloss. To compute the Lesk similarity (Banerjee & Pedersen 2002) between A and B, a scoring function based on the overlap of words in their individual glosses is used.

Leacock and Chodorow (LCH): To measure similarity between two concepts A and B, Leacock & Chodorow (1998) compute the shortest path through hypernymy relation between them under the constraint that there exists such a path. The final value is computed by scaling the path length by the overall taxonomy depth (D).

$$sim_{LCH}(A, B) = -\log \left(\frac{len(A, B)}{2D} \right)$$

6 Experimental setup

Word sense disambiguation of WordNet glosses is carried out using the WSD engine by Zhong & Ng (2010). It is an all-words generic WSD engine with an accuracy of 82% on standard WSD corpus. We use WordNet::Similarity 2.05 package by Pedersen et al. (2004) for computing the similarity by other metric scores mentioned in this paper. We use Pearson correlation coefficient to find the inter-annotator agreement.

For evaluation based on synset replacement using similarity metrics, we use the dataset provided by Balamurali et al. (2011). The experiments are performed using C-SVM (linear kernel with default parameters³), using bag-of-synsets as features, available as a part of LibSVM⁴ package. All classification results reported are average of five-fold cross-validation accuracies.

To evaluate the result, we use accuracy, recall and precision as the metrics. Classification accuracy defines the ratio of the number of true instances to the total number of instances. Recall is calculated as a ratio of the true instances found to the total number of false positives and true positives. Precision is defined as the number of true instances divided by the number of true positives and false negatives. Positive Precision (PP) and

Positive Recall (PR) are precision and recall for positive documents while Negative Precision (NP) and Negative Recall (NR) are precision and recall for negative documents.

7 Results and discussion

7.1 Sentiment as a parameter for finding similarity

Table 3 shows correlation scores between two annotators for different annotation strategies. Apart from the correlation of the complete word pairs, it also shows POSwise correlation. From the results, it is clearly evident that similarity is best captured when sentiment is also included. The annotation strategy involving a combination of meaning and sentiment has a better correlation among annotators(0.799) than the one which considers meaning alone. This verifies the hypothesis that taking sentiment content of words being compared is beneficial in assessing the similarity between them.

A POSwise analysis of Table 3 suggests that apart from the Noun category, all other categories have a better correlation among annotators in assessing the sentiment based similarity annotation strategy. This may be due to the fact that in case of word pairs belonging to the Noun category, sentiment does not play much role. The highest correlation is seen for Verbs. In case of Adjectives, annotators have a fairly high correlation. Since most of the Adjectives are sentiment bearing words, annotators might have found it easier to compare them. In summary, a similarity metric which incorporates sentiment may be more beneficial to POS categories other than Nouns.

Table 4 shows the correlation between scores obtained using different similarity metrics with the scores obtained from gold standard dataset of an annotator. The correlation with respect to different POS categories is also shown. NA in some of the columns represent word pairs , belonging to those POS categories, which cannot be handled using the similarity metric concerned. For example, metrics like LIN and LCH cannot handle POS categories other than Nouns and Verbs. SenSim

³C=0.0,epsilon=0.0010

⁴<http://www.csie.ntu.edu.tw/~cjlin/libsvm>

Metric Used	Overall	NOUN	VERB	ADJECTIVES	ADVERBS
LESK	0.22	0.51	-0.91	0.19	0.37
LIN	0.27	0.24	0.00	NA	NA
LCH	0.36	0.34	0.44	NA	NA
SenSim (SD)	0.46	0.73	0.55	0.08	0.76
SenSim (TD)	0.50	0.62	0.48	0.06	0.54
SenSim (SM)	0.45	0.73	0.55	0.08	0.59
SenSim (TM)	0.48	0.62	0.48	0.06	0.78

Table 4: Pearson Correlation(r) of various metrics with Gold standard data

Metric used	Accuracy(%)	PP	NP	PR	NR
Baseline	89.10	91.50	87.07	85.18	91.24
LSK	89.36	91.57	87.46	85.68	91.25
LIN	89.27	91.24	87.61	85.85	90.90
LCH	89.64	90.48	88.86	86.47	89.63
SenSim (SD)	89.95	91.39	88.65	87.11	90.93
SenSim (TD)	90.06	92.01	88.38	86.67	91.58
SenSim (SM)	90.11	91.68	88.69	86.97	91.23
SenSim (TM)	90.17	91.81	88.71	87.09	91.36

Table 5: Classification results of synset replacement experiment using different similarity metrics; PP-Positive Precision (%), NP-Negative Precision(%), PR-Positive Recall (%), NR-Negative Recall (%)

has a better correlation with the gold standard data than other metrics. In fact, all variants of SenSim function better than the existing similarity measurement techniques used in this paper. Among different variants, SenSim using TD based scoring function performs the best. It has a correlation score of .50 whereas the nearest correlation among the other metrics is obtained using LCH (.36). Moreover, in case of all POS categories barring Adjectives, SenSim metric has a better correlation than the rest of the metrics with gold standard data.

Although not provided in the table, reader should note that the metrics used in this study could not score all the word pairs created for this study. For example, out of 48 word pairs, SenSim could mark only 34 word pairs and among other metrics, the best count is provided by Lesk with 17 word pairs. The correlation scores shown for each metric is with respect to the word pairs that had some values for comparison with the gold standard data. Thus in terms of coverage also, SenSim performs better than the metrics used in this study.

7.2 Effect of SenSim on synset replacement strategy

Table 5 show the results of document level sentiment classification using synset replacement strat-

egy based on similarity metrics. The results of the classifier trained on synsets alone, without any replacement, is taken as the baseline.

SenSim(TM) variant obtains the best classification accuracy among the various metrics analyzed. It achieves an accuracy of 90.17%. Even though the improvement is marginal, reader must note that no complex features are used for training this classifier. All variants of SenSim, except for SenSim(SD) achieve a 90% classification accuracy. Compared to the baseline, other WordNet metrics also have better classification accuracies.

Classification using SenSim (TM) metric has the highest positive precision. Same is the case with negative recall, it has a negative recall of 91.58%. The SenSim (SD) variant has the highest positive recall. In general, it can be seen that the positive class's performance has improved by using SenSim metric for synset replacement.

8 Conclusion

In this paper, we proposed that sentiment content can aid in similarity measurement. We verified this hypothesis by taking the correlation between annotators using different annotation strategies. Annotator correlation for the strategy involving sentiment as an additional parameter for similarity measurement was higher than the one

which involved just semantic similarity. Based on this hypothesis, we introduced a new similarity metric, SenSim, which accounts for the sentiment content of the words being compared. An intrinsic evaluation of the metric with human annotated word pairs for similarity showed higher correlations than the popular WordNet based similarity metrics. We also carried out an extrinsic evaluation of SenSim on synset replacement strategy for the mitigation of unknown feature problem in supervised classification. Our results suggest that apart from the overall improvement of sentiment classification accuracy, SenSim improves the classification performance of the positive-class-documents .

Acknowledgments

We thank Jaya Saraswati and Rajita Shukla from CFILT Lab, IIT Bombay for annotating the dataset used for this work. We also thank Mitesh Khapra and Salil Joshi, IIT Bombay for providing us with the WSD engine.

References

- Balamurali, A., Joshi, A. & Bhattacharyya, P. (2011), Harnessing wordnet senses for supervised sentiment classification, *in* ‘Proc. of EMNLP-2011’.
- Banerjee, S. & Pedersen, T. (2002), An adapted lesk algorithm for word sense disambiguation using wordnet, *in* ‘Proc. of CICLING-02’.
- Banerjee, S. & Pedersen, T. (2003), Extended gloss overlaps as a measure of semantic relatedness, *in* ‘Proc. of IJCAI-03’.
- Esuli, A. & Sebastiani, F. (2006), SentiWordNet: A publicly available lexical resource for opinion mining, *in* ‘Proceedings of LREC-06’, Genova, IT.
- Grishman, R. (2001), Adaptive information extraction and sublanguage analysis, *in* ‘Proc. of IJCAI-01’.
- Hirst, G. & St-Onge, D. (1997), ‘Lexical chains as representation of context for the detection and correction malapropisms’.
- Jiang, J. J. & Conrath, D. W. (1997), Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, *in* ‘Proc. of ROCLING X’.
- Leacock, C. & Chodorow, M. (1998), Combining local context with wordnet similarity for word sense identification, *in* ‘WordNet: A Lexical Reference System and its Application’.
- Leacock, C., Miller, G. A. & Chodorow, M. (1998), ‘Using corpus statistics and wordnet relations for sense identification’, *Comput. Linguist.* **24**.
- Lin, D. (1998), An information-theoretic definition of similarity, *in* ‘Proc. of ICML ’98’.
- Partington, A. (2004), ‘Utterly content in each others company semantic prosody and semantic preference’, *International Journal of Corpus Linguistics* **9**(1).
- Patwardhan, S. (2003), Incorporating dictionary and corpus information into a context vector measure of semantic relatedness, Master’s thesis, University of Minnesota, Duluth.
- Pedersen, T., Patwardhan, S. & Michelizzi, J. (2004), Wordnet::similarity: measuring the relatedness of concepts, *in* ‘Demonstration Papers at HLT-NAACL’04’.
- Rada, R., Mili, H., Bicknell, E. & Blettner, M. (1989), ‘Development and application of a metric on semantic nets’, *IEEE Transactions on Systems Management and Cybernetics* **19**(1).
- Resnik, P. (1995a), Disambiguating noun groupings with respect to Wordnet senses, *in* ‘Proceedings of the Third Workshop on Very Large Corpora’, Somerset, New Jersey.
- Resnik, P. (1995b), Using information content to evaluate semantic similarity in a taxonomy, *in* ‘Proc. of IJCAI-95’.
- Richardson, R., Smeaton, A. F. & Murphy, J. (1994), Using wordnet as a knowledge base for measuring semantic similarity between words, Technical report, Proc. of AICS-94.
- Sinclair, J. (2004), *Trust the Text: Language, Corpus and Discourse*, Routledge.
- Wan, S. & Angryk, R. A. (2007), Measuring semantic similarity using wordnet-based context vectors., *in* ‘Proc. of SMC’07’.
- Wu, Z. & Palmer, M. (1994), Verb semantics and lexical selection, *in* ‘Proc. of ACL-94’, New Mexico State University, Las Cruces, New Mexico.
- Zhong, Z. & Ng, H. T. (2010), It makes sense: A wide-coverage word sense disambiguation system for free text., *in* ‘ACL (System Demonstrations)’10’.

Toward Building a Large-Scale Arabic Sentiment Lexicon

Muhammad Abdul-Mageed

Department of Linguistics &

School of Library & Information Science,
Indiana University,
Bloomington, USA,
mabdulma@indiana.edu

Abstract

Subjectivity and sentiment analysis (SSA) depends heavily on the availability of polarity lexicons where entries are tagged with semantic orientation (e.g., *positive*, *negative*, and *neutral* tags). Although there has been a swelling interest in building English language polarity lexicons, there has not been such major efforts for Morphologically-Rich Languages (MRL) such as Arabic. In this study, we report efforts to expand a manually-built polarity lexicon of Modern Standard Arabic (MSA) using existing English polarity lexica exploiting a simple machine translation procedure. Using glosses from our manually-tagged lexicon as gold standard, we evaluate the expanded lexicon and point out directions for further improvements. Finally, we use our gold tags to evaluate the coverage and quality of a sub-segment of SentiWordNet 3.0, a widely-used polarity lexicon.

1 Introduction

The area of *Subjectivity and Sentiment Analysis* (SSA) has recently been receiving a lot of interest both within the academia and the industry. In natural language, the term *subjectivity* refers to expression of opinions, evaluations, feelings, and speculations (Banfield, 1982; Wiebe, 1994) and thus incorporates *sentiment*. The process of *subjectivity classification* refers to the task of classifying texts into either *objective* (e.g., *Al-Asad is still the president of Syria*) or *subjective* (e.g., *Mubarak, the hateful dictator, stepped down*). Subjective text is further classified with *sentiment* or *polarity*. For sentiment classification, the task refers to identifying whether the subjective text is *positive* (e.g., *Wow! It's such an amazing camera!*), *negative* (e.g., *I really hate this restaurant!*), *neutral* (e.g., *I believe the new iPhone will be released soon.*), or, sometimes, *mixed* (e.g., *Well, it's efficient, but the design is ugly!*) texts.

Mona T. Diab

Center for Computational

Learning Systems,

Columbia University, NYC, USA,
mdiab@ccls.columbia.edu

SSA heavily depends on a knowledge of the behavior of language items within their corresponding *lexical fields*. A lexical field is the set of lexical items that cover a specific concept (Lyons, 1977). For example, the field of anger terms may contain units as “rage,” “fume,” “seethe,” “boil over,” and “look daggers”. Expressions are also said to have a *semantic orientation* (SO) (also called *prior polarity*). The SO of an expression indicates the direction the expression deviates from the norm for its lexical field: It is an evaluative characteristic (Battistella, 1990) of the meaning of the word that restricts its usage to an appropriate context (Hatzivassiloglou and McKeown, 1997; Hatzivassiloglou and Wiebe, 2000). For example, in a field describing a lodging place, whereas the words “full” and “empty” may have a neutral SO, the words “fascinating” and “admirable” have a positive SO, and the words “shocking” and “detestable” have a negative SO. Due to the usefulness of prior polarity for the task of SSA, several attempts have been made to build prior polarity lexica, especially English language ones. To date, there are no attempts to build a large scale Arabic polarity lexicon for use with SSA. In the current study, we report efforts to build a wide coverage polarity lexicon of Arabic, both manually and automatically. In the process, we leverage manually- and automatically-developed English lexica and evaluate SentiWordNet (SWN3) (Esuli and Sebastiani, 2005), a widely-used lexical database for SSA. Finally, we point out directions for future research.

2 Building an Arabic Polarity Lexicon

2.1 PATB Adjectives (Sifaat)¹

We introduce *Sifaat*, a manually created lexicon of 3325 Arabic adjectives labeled with one of the following tags {*positive* (*Pos*), *negative* (*Neg*), *neu-*

¹Sifaat or “صفات” in Arabic script is Arabic for “adjectives.”

tral (*Obj*)}. The adjectives in Sifaat pertain to the newswire domain and were extracted from the first four parts of the Penn Arabic Treebank (PATB) (Maamouri et al., 2004; Abdul-Mageed and Diab, 2011). The lexicon was tagged with two college-educated independently and cases of differences were settled by the two annotators after adjudication. Annotators were instructed not to use their personal experiences/backgrounds when labeling the entries, but what they believe would be the orientation of a general use of each term. Annotators were also instructed to assign what they think is the majority orientation of a term, in cases where a term may have both positive and negative use depending on context. The annotators were also exposed to a linguistic background about the concepts of subjectivity and sentiment and the task of annotating data for these concepts as the task is explained in (Abdul-Mageed and Diab, 2011). Although labeling lexica outside context is different from labeling text units (e.g., sentences) as explained in (Abdul-Mageed and Diab, 2011) (e.g., speech act, good vs. bad news, annotator’s background knowledge) are still relevant. Examples of the adjectives labeled as Pos are “مشوق” “splendid” and “بطولي” “heroic”. Examples of entries assigned a Neg tag are “مشؤوم” “inauspicious” and “محرضي” “provocative”. In addition to the polarity labels, Sifaat is enriched with English glosses for its entries. Sifaat was successfully used for enhancing Arabic SSA as reported in (Abdul-Mageed et al., 2011a), with about 6% *F*-measure improvement for subjectivity classification and more than 40% *F*-measure improvement for sentiment classification. These results show the utility and need for a resource such as Sifaat for Arabic SSA.

2.2 Leveraging Available English Lexica

Recognizing the need for a wide coverage lexical resource that covers multiple genres in Arabic, We experiment with fast automatic acquisition of an Arabic polarity lexicon using three existing English lexica. We introduce each of these lexica below and then we introduce our method.

2.2.1 WordNet and SentiWordNet (SWN)

WordNet (Miller, 1995; Fellbaum, 1998) is a large lexical database of English where nouns, verbs, adjectives, and adverbs are grouped into sets of synonyms (synsets), each expressing a distinct concept. The synsets are interlinked by means of conceptual-semantic and lexical relations.

(Esuli and Sebastiani, 2005) developed a method for determining the orientation of a term based on the classification of its glosses as derived from WordNet. They use terms derived from WordNet to build *SentiWordNet*, a lexical database with polarity scores assigned to each synset. Later, (Baccianella et al., 2010) introduce a new version of the database, SentiWordNet 3.0 (SWN3), where they use a random-walk step that consists of viewing WordNet 3.0 as a graph, and running an iterative, “random-walk” process in which the numerical scores, *Pos(s)* and *Neg(s)* (and, hence, *Obj(s)*), starting from those determined in the previous step, possibly change at each iteration. The random-walk step terminates when the iterative process has converged. They add that the basic intuition is that, if most of the terms that are being used to define a given term are positive (resp., negative), then there is a high probability that the term being defined is positive (resp., negative) too. Positivity and negativity are thus seen as “flowing through the graph,” from the terms used in the definitions to the terms being defined. SentiWordNet is composed of more than 117,000 synsets from which we acquire 192,493 unique entries.

2.2.2 Youtube Lexicon (YT)

(Abdul-Mageed et al., 2011b) used Google’s Youtube Data API to crawl all comments on 1000 Youtube videos using the query “obama health care”. They refer to the 229,177-comments resulting corpus as *Youtube Health Corpus [YuHC]*. After reducing all repeated letters of frequency > 2 to only 2 (e.g., the word *coool* is reduced to *cool*), they extracted the top 29,991 words² and manually labeled them with semantic orientation tags. Each word was given a label of the set {*Pos*, *Neg*, *Obj*}.

²Extracted words were of frequency of 3 or more.

Lexicon	# of entries
Sifaat-Pos	617
Sifaat-Neg	550
Sifaat-Obj	2,158
SWN-Pos	17,323
SWN-Neg	18,563
SWN-Obj	156,607
YT-Pos	3,768
YT-Neg	6,224
YT-Obj	19,999
GI-Pos	1,636
GI-Neg	2,007
All-Pos	23,344
All-Neg	27,344
All-Obj	178,764
All	229,452

Table 1: Acquired lexicon

2.2.3 General Inquirer (GI)

The General Inquirer (GI) lexicon (Stone et al., 1966) is part of the GI system, a content analysis program that exploits terms manually classified on a large number of categories. The lexicon contains a total of 11788 terms, 1636 of them are labeled as Positive and 2007 are labeled as Negative, whereas the remaining items, not assigned either negative or positive tags, maybe considered Objective. The GI lexicon has been widely used in SSA, e.g. (Kamps et al., 2004; Esuli and Sebastiani, 2005; Turney and Littman, 2003).

3 Approach

3.1 Expanding Sifaat

We use Google’s translation API to render all expressions in SWN3, YT, and GI lexica into Arabic. Table 1 shows the number of entries in the various lexica and the total number of entries per category.³ As can be calculated based on Table 1, the acquired lexicon has 14613 entries that do not exist in SWN3.

We thus acquire an expanded lexicon totaling 229,452 entries. For each entry, in addition to the Arabic form, we store the English gloss, the Buckwalter transliteration, and a sentiment tag from the

Arabic	Gloss	Buckwalter	Tag
إصلاح	reform	ISIAH	Pos
صديق	friend	Sdyq	Pos
كرامة	dignity	krAmp	Pos
المؤمنين	believers	AlmWmnyn	Pos
سيف	fucking	sxyf	Neg
أحمق	idiot	OHmq	Neg
كسل	idleness	ksl	Neg
خسيس	contemptible	xsys	Neg
رصيف	pavement	rSyf	Obj
الوعد	promisee	AlmwEwd	Obj
أنفي	nasal	Onfy	Obj
فلسطيني	Palestinian	fIsTny	Obj

Table 2: Example entries from expanded lexicon

set $\{Pos, Neg, Obj\}$. Table 2 shows a number of entries from the expanded lexicon.

We then use the English glosses from Sifaat to evaluate the adjectives of SWN3. We first average the scores assigned to the various senses of each SWN3 adjective. Then, we search SWN3 for each Sifaat adjective gloss and if found we compare the tags assigned to it in each lexicon. Using this procedure we map the Sifaat adjectives to their counterparts in the SWN3 pool of 29,816 adjectives. The tags assigned agree with a Kappa (k) = 0.284. This indicates a fair agreement (Landis and Koch, 1977). In order to gain higher agreement rates, we experimented with two thresholds of scores. Using the thresholds, we retrieve two sub-SWN3 databases, as follows:

SentiWN_Lex: This lexicon is composed of all Pos ($N= 1,949$) and Neg ($N= 2,580$) entries with a score > 0.25 plus all Obj entries ($N= 5,243$). The list thus totals 9,772 entries.

SentiWN_Strong_Lex: This lexicon is composed of all Pos ($N= 624$) and Neg ($N= 1,156$) entries with a score > 0.50 as well as all the Obj entries mentioned above (i.e., $N= 5,243$). The list totals 7,023 entries.

We find that the higher the threshold, the better the agreement we acquire. Thus, with the > 0.25 threshold, we get a higher (although still ‘fair’) agreement (with a Kappa (k) = 0.346). With the higher thresh-

³We do not include GI-Obj.

old of > 0.50 , a 'moderate' agreement (Landis and Koch, 1977) with a higher Kappa (k) = 0.426 is achieved. However, the higher threshold comes at the cost of lower coverage (i.e., 7,023 SWN3 adjectives with the > 0.50 threshold vs. the whole 29,816 SWN3 adjective pool).

3.2 Comparing SWN with YT

Unlike SWN, YT was manually built and is domain-specific. We compared the two lexica and found that 13,945 YT entries do not exist in SWN3. The big number is due partly to the fact that the YT lexicon includes various word forms for each word type (e.g., it has the token "protesters" along with the type "protest"), it is also due to lack of certain slang terms (e.g., "bruh," "dammit") and named entities (e.g., "Obama," "Mugabe") in SWN. This shows that even though SWN3 has wide coverage, it is not inclusive of some of the expressions widely used in the language of social media. In addition, we observe that some commonly-used expressions (e.g., "suck," "lol") that bear strong sentiment in YT are tagged as Obj in SWN3, which, again, shows that SWN3 lacks some of the nuances characteristic of social media language use.

4 Related Work

In the SSA literature, a seed list of items with a clear prior polarity has been used to identify the polarity of additional items and hence build a prior polarity lexicon. Such a lexicon is then used to enhance the task of automatically detecting the subjectivity and/or sentiment of texts at levels beyond the phrase (e.g., sentences, paragraphs, documents). Several approaches have been proposed for learning an expression's prior polarity. Some researchers, e.g., (Dave et al., 2003; Kim and Hovy, 2004; Mullen and Collier, 2004) have successfully used WordNet (Miller, 1995; Fellbaum, 1998) for, e.g., retrieving synonyms and antonyms of elements in the seed set.

(Hatzivassiloglou and McKeown, 1997) report efforts to learn the semantic orientation of adjectives from a corpus. They maintain that *conjunctions* between adjectives are specifically useful, since they impose constraints on the semantic orientation of their arguments. Another observation they make is that conjoined adjectives that have related forms

(e.g., "adequate-inadequate") almost always have different semantic orientation. Based on these observations, these authors adopt an unsupervised learning algorithm to infer the orientation of adjectives and report an accuracy of 92.37%. One advantage of (Hatzivassiloglou and McKeown, 1997)'s method is that it is unsupervised, and so there is no need to manually annotate data. However, the method assumes the existence of a huge POS-tagged corpus, which is challenging due to the potential unavailability of either huge corpora or high-performance POS taggers, or both, for some languages.

(Turney, 2002) presents an unsupervised algorithm for learning the prior polarity of adjectives or adverbs in phrases in a POS-tagged corpus. He measures the orientation of the phrases using a statistical measure based on Pointwise Mutual Information (PMI) (Church and Hanks, 1990). He calculates the semantic orientation (SO) of a phrase by comparing its similarity to the word "excellent" to its similarity to the word "poor." (Turney, 2002) reports a classification accuracy of 74.39% across the four domains of his data. (Turney, 2002)'s algorithm is conceptually simple, but is constrained in the sense that it is dependent on results retrieved from a search engine with specific settings (i.e., use of the NEAR operator) that no longer exist.

5 Conclusion

In this study, we report efforts to expand a manually-built polarity lexicon of MSA using several existing English polarity lexica. We also exploit glosses from the manually-tagged lexicon to evaluate both the expanded lexicon and a sub-segment of SWN3. We find some problems with both the coverage and quality of some SWN3 entries and believe that incrementing it with social-media lexica is a useful measure. Although we have not tested the utility of the expanded lexicon in an SSA system, we believe it will be useful. We plan to further expand the lexicon we acquired by manually-labeling entries extracted from an Arabic social media corpus.

References

- M. Abdul-Mageed and M. Diab. 2011. Subjectivity and sentiment annotation of modern standard arabic

- newswire. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 110–118, Portland, Oregon, USA, June. Association for Computational Linguistics.
- M. Abdul-Mageed, M. Diab, and M. Korayem. 2011a. Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Muhammad Abdul-Mageed, Mohammed Korayem, and Ahmed YoussefAgha. 2011b. yes we can?: Subjectivity annotation and tagging for the health domain. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 666–671, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.
- S. Baccianella, A. Esuli, and F. Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Seventh conference on International Language Resources and Evaluation, Malta*. Retrieved May, volume 25, page 2010.
- A. Banfield. 1982. *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Routledge & Kegan Paul, Boston.
- E.L. Battistella. 1990. *Markedness: the evaluative superstructure of language*. State Univ of New York Pr.
- K.W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.
- K. Dave, S. Lawrence, and D.M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM.
- A. Esuli and F. Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 617–624. ACM.
- C. Fellbaum. 1998. *WordNet: An electronic lexical database*. The MIT press.
- V. Hatzivassiloglou and K. McKeown. 1997. Predicting the semantic orientation of adjectives. In *ACL-EACL*, pages 174–181.
- V. Hatzivassiloglou and J. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *International Conference on Computational Linguistics, (COLING-2000)*.
- J. Kamps, MJ Marx, R.J. Mokken, and M. De Rijke. 2004. Using wordnet to measure semantic orientations of adjectives.
- S. Kim and E. Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373.
- J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.
- J. Lyons. 1977. Semantics. 2 vols.
- M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. 2004. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109.
- G.A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- T. Mullen and N. Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP) Conference*, pages 412–418, Barcelona, Spain.
- P.J. Stone, D.C. Dunphy, and M.S. Smith. 1966. The general inquirer: A computer approach to content analysis.
- P. Turney and M.L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. In *ACM Transactions on Information Systems (TOIS)*.
- P. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*.
- J. Wiebe. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

WordNet Atlas: a web application for visualizing WordNet as a zoomable map

Matteo Abrate, Clara Bacci, Andrea Marchetti and Maurizio Tesconi

Istituto di Informatica e Telematica (IIT) - CNR

Via G. Moruzzi, 1 - Pisa, Italy

{matteo.abrate, clara.bacci, andrea.marchetti,
maurizio.tesconi}@iit.cnr.it

Abstract

The English WordNet is a lexical database containing more than 206,900 word-concept pairs and more than 377,500 semantic and lexical links. Trying to visualize them all at once as a node-link diagram results in a representation which can be too big and complex for the user to grasp, and which cannot be easily processed by current web technologies.

We propose a visualization technique based on the concept of *semantic zooming*, usually found in popular web map applications. This technique makes it feasible to handle large and complex graphs, and display them in an interactive representation. By zooming, it is possible to switch from an overview visualization, quickly and clearly presenting the global characteristics of the structure, to a detailed one, displaying a classic node-link diagram.

WordNet Atlas is a web application that leverages this method to aid the user in the exploration of the WordNet data set, from its global taxonomy structure to the detail of words and synonym sets.

1 Introduction

A big variety of data sources often contain a massive amount of interlinked elements, forming very large graph structures: Telecommunication traffic, social networks and the World Wide Web are all prominent examples of sources that could yield graphs having more than 1 million edges (Abello et al., 2002; Cui, 2011). Because of their size and topology, such structures could be very difficult to explore, even for expert users. Getting an overview of or finding detailed information in them can prove to be complicated, and a poorly de-

signed interface could be of little help if it makes the user feel overwhelmed or disoriented.

The information visualization community agrees that providing a visual representation of graphs is of a great importance for end users. Visual depictions exploit human visual processing to reduce the cognitive load of many tasks that require understanding of global or local structure of data (Munzner, 2000; Ware, 2004; Zhang, 1991).

To obtain such representations, the data must be preprocessed and interpreted. For example, input data can be simplified by highlighting or hiding some features, considered respectively more or less interesting for a given task. The data can also be enriched by derived features.

Size is a major challenge in graph visualization, leading to problems like algorithm complexity, display cluttering, poor readability and difficulties in navigation (Cui, 2011). The aforementioned issues are also found within the context of visualization of smaller-scale graphs, where the size is still a relevant issue (100,000 edges or more). Structures of that size could be obtained from document collections such as Wikipedia, Linked Data¹ datasets like DBpedia or GeoNames, lexical databases like WordNet.

The English WordNet version 3.0 (Miller, 1995; Fellbaum, 1998) contains 155,287 words linked to 117,659 sets of synonyms (synsets) each representing a concept, defining 206,941 word-synset pairs². Synsets are grouped by part-of-speech in four sub-structures: nouns, verbs, adjectives, and adverbs. Moreover, words and synonym sets are interconnected by 285,348 semantic links (sem-links) and 92,231 lexical links (lexlinks) of 28 different types. A complete visual representation of WordNet using a standard node-link diagram would result in a particularly heavy graphic to pro-

¹<http://linkeddata.org>

²<http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html>

cess and display, and would probably be too big and complex for users to understand.

Current web technologies cannot cope with such a number of elements without special techniques. These techniques should somehow reduce the amount of information sent to the client and rendered by the user agent, without excessively compromising the meaning of the graphic or worsening the quality of user interaction. Moreover, without choosing a good layout, the WordNet graph could easily render in a large, intricate and unorganized diagram.

This paper describes an interactive map³ based on a modified radial tree layout. It lets the user navigate the entire structure of the WordNet graph of noun synsets and makes him capable to have a general overview, as well as to concentrate on details on demand (Shneiderman, 1996). We believe that this visualization adds usability and usefulness to WordNet, because it helps the user to get oriented in the large amount of data it contains.

1.1 Outline of the paper

In section 1, we discussed the main issues related to the visualization of graphs obtained from large data sets, such as WordNet.

In section 2, we provide a background reference for general visualization concepts and methods, ending with a brief survey of the state of the art of web interfaces for browsing WordNet (subsection 2.2).

In section 3, we describe the proposed solution, along with a report of the analysis of WordNet’s structure (subsection 3.1) and implementation details (subsection 3.2).

In section 4, we present our conclusions and hint at possible directions for further research.

2 Background

2.1 Graph visualization concepts

It is known that our perception of things is mostly due to vision: we acquire more information through sight than through all the other senses combined (Ware, 2004). This is why a good visual representation can be a powerful interface between large amounts of data and our cognitive system, helping us in analysis and decision-making tasks. It has been pointed out by Ware (2004) that a good visualization has five big advantages:

³The web application will be publicly available at <http://wafi.iit.cnr.it/wordnetatlas>

- It gives us the ability to comprehend large amounts of data;
- It lets us perceive properties that were non evident before;
- It makes problems with the data immediately apparent;
- It facilitates us in understanding both large-scale and small-scale features of the data;
- It facilitates hypothesis formation about the data.

To best exploit such advantages, visualizations can be carefully designed in a way that conveys meaning while keeping aesthetics into account. This kind of visualizations are often referred to as *infographics* (information graphics).

In his survey, Cui (2011) describes the main issues behind graph visualization, along with layout techniques, like node-link and space division, and interaction styles, such as filtering and *semantic zooming*. The latter technique improves a typical *zoom and pan* approach by introducing the concept of Level Of Detail (LOD), thus suggesting to display the graph diagram in a *Zooming User Interface*.

Zooming User Interfaces (ZUI) are a kind of graphical interfaces that give the user the ability to zoom and pan an environment. Within this environment, visual representations of the data are displayed with different Level Of Detail according to the zoom level. Very popular and successful services that embrace this approach are web mapping applications such as Google Maps⁴ or Bing Live Maps⁵, where the user could quickly zoom out for an overview of the whole planet and zoom in for observing the details of a particular place. It is our opinion that the semantic zooming approach could prove to be valid outside the context of digital cartography, and in particular within the context of visualization of big graphs such as WordNet.

Many tools have been developed to help researchers to visualize graphs, even of big size. These tools can be very useful to analyze a graph to envision a meaningful graphical representation to base an infographic upon.

Gephi⁶ is an interactive visualization and exploration platform for various kinds of networks

⁴<http://maps.google.com>

⁵<http://www.bing.com/maps/>

⁶<http://gephi.org/>

and complex systems, dynamic and hierarchical graphs up to 50 thousand nodes and 500 thousand edges. Tulip⁷ (Auber, 2003) is a framework to interactively visualize graphs up to 1 million elements, that makes several graph drawing, clustering and visual mapping algorithms available.

Among others, Visone⁸ is a research project that designs and implement an interactive visualization software for social network analysis. Cytoscape⁹ is an open source platform designed for biological research, now used to display and analyze different kind of networks. NodeXL¹⁰ is a template for Microsoft Excel that integrates a graph visualization and analysis tool into the application.

2.2 Web interfaces to WordNet

There are many different interactive visualizations of the WordNet graph that complement the traditional WordNet Search¹¹. They often show words of different part-of-speech, synsets and senses as graph nodes, and word-synset links, semlinks and lexlinks as edges. They show a small subgraph of WordNet, created on the fly. This graph is dynamically adjusted as the user chooses to navigate to other nodes, selected by clicking or by issuing a word search.

WordVis¹², Synonym¹³ and Javascript Visual-Wordnet¹⁴ show the graph from the vantage point of a selected word or synset. It centers the selected node in the view, showing details about it and deleting the nodes that become not directly connected to it. WordNet Editor¹⁵, Visuwords¹⁶ and the visualization tool¹⁷ described in the work by Kamps (2002) actually allow a deeper exploration of the graph: the user can click on a node and the connected nodes are added to the view, without deleting the old ones. The disadvantage of these approaches is the progressive performance deterioration as nodes and links are added. Treebolic¹⁸ uses a different kind of graph visualization compared with the previous ones: the nodes po-

sition is computed as soon as the graph is drawn, they remain still (they don't adjust their position to better fit the screen), and a *fisheye view* (Furnas, 1986) is used to focus on the details. Even in this case, though, the graph is partial and centered in a specific word sense. Collins (2007) uses a fisheye technique too, giving detail as well as a good level of overview. The showed graph is never complete, though, and it is always rooted in the selected node.

3 WordNet Atlas

The potentially huge amount of nodes and edges of a graph could be organized in a meaningful spatial layout, alongside artificially added spatial features such as colored regions or placemarks, possibly derived from graph data or measures. The resulting visualization will be a mix between a classic node-link diagram and a so-called infographic, a graphic way to present information that abstracts and represents many qualitative and quantitative aspects of a subject in a carefully studied design.

To make the visualization and exploration of the WordNet graph more convenient for the users, we tried to make it easier for them to get an orientation and remember where the nodes are in the map, and consequently to enhance their ability to find them again (revisitation). This is achieved by drawing an interactive graphic that gives a stable representation of the synset nodes (they are always drawn in the same position) and offers the users an overview of the entire data set, along with some added *landmarks*, useful to improve their own orientation. Since the diagram is intended for presentation and for a better fruition mostly by non experienced WordNet users, some simplification in data has been made. Therefore, the map is far from being a perfect representation of the English WordNet lexical database.

In WordNet Atlas, we choose to show two different visualizations at once, one best suited for exploring the graph of synsets, and one for navigating among words (see figure 1).

The first is an interactive map of a graph, representing all the WordNet noun synsets arranged in a special circular layout. At the center of the map is the most generic node of the noun synset taxonomy, the one containing the word "entity"¹⁹, serving as the root of a radial tree layout.

To further characterize the representation, we

⁷<http://tulip.labri.fr/TulipDrupal/>

⁸<http://www.visone.info/>

⁹<http://www.cytoscape.org/>

¹⁰<http://nodexl.codeplex.com/>

¹¹<http://wordnetweb.princeton.edu/perl/webwn>

¹²<http://wordvis.com/>

¹³<http://code.google.com/p/synonym/>

¹⁴<http://kylescholz.com/projects/wordnet/>

¹⁵<http://wordventure.eti.pg.gda.pl/wne.html>

¹⁶<http://www.visuwords.com/>

¹⁷<http://staff.science.uva.nl/kamps/wordnet/>

¹⁸<http://id.asianwordnet.org/visualize/treebolic/>

¹⁹<http://wordnet.princeton.edu>

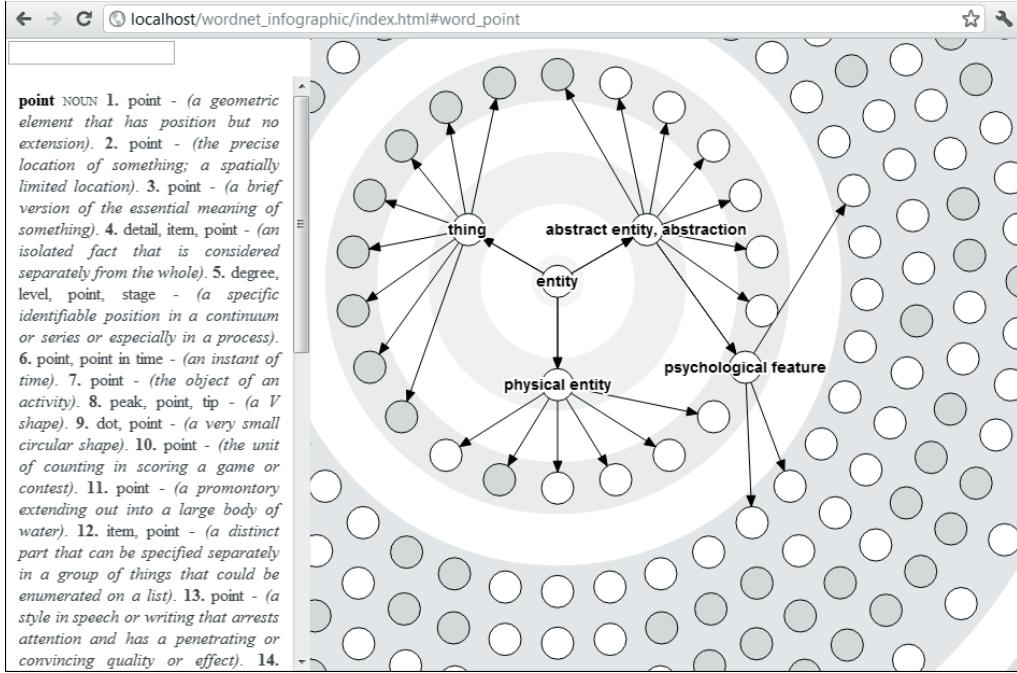


Figure 1: Screenshot of WordNet Atlas prototype interface.

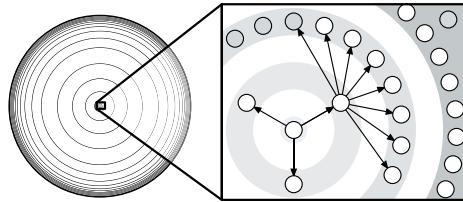


Figure 2: Different Levels Of Detail: at minimum zoom (left), the whole graphic of the noun synsets, divided into rings, and at a greater zoom (right), the central region, in which the classic node-link diagram is visible.

decided to display each level of the tree as a colored ring underlying the representation of nodes. This should also have the effect of enhancing the graph memorability by leveraging the user's spatial memory (Ghani and Elmquist, 2011).

To address both the user's inability to cope with large amounts of displayed symbols and the technical problems of performance of current web technologies, we decided to follow a *semantic zooming* approach.

At the minimum Level Of Detail, the entire

graph is shown as a circle made of concentric rings, representing the depth level of the taxonomy tree. The ring thickness depends on how many synset are present at that depth level. Rings play the role of *spatial landmarks* (Ghani and Elmquist, 2011) while giving an overview of the taxonomy tree structure.

By zooming in, the user can load the actual synset nodes grouped in the rings, shown as white circles. Rings are still displayed as colored regions under the graph, following the technique of *substrate encoding* (Ghani and Elmquist, 2011). The user can use this visual information to more easily recognize the location of the map he or she is looking at.

By interacting with the synset, the user could reach the maximum level of detail, obtaining fine grained information such as the synset's definition, its set of synonym words and semantic links (edges) that connect it with other synsets in different points of the map.

We developed a second visualization, complementary to the map: an hypertextual dictionary shown at the left side of the interface. The user can search for a word entry by using the provided search box, or scroll the whole dictionary by drag-

ging a scroll bar. Like a paper dictionary, each entry displays a word, its part-of-speech and a list of numbered word senses. For each sense, a list of synonym words and a definition are displayed. Not unlike the map, we decided to give the user an overview of the amount of available information in the word database by giving them the illusion that the whole dictionary was loaded in a scrollable section of the interface.

It is possible to jump from a word to another within the dictionary by clicking it in the synonym lists or in the map, or to jump to a synset within the map by clicking its definition. This behavior integrates seamlessly with the standard user agent interface, so the user is able to exploit the familiar navigation buttons, the browser history and bookmarks to navigate words and synsets.

3.1 Analysis of the WordNet database

In order to find useful properties to obtain the described visualization, we analyzed an SQL version of the WordNet 3.0 database.

We chose the synsets nouns as graph nodes, trying to identify a hierarchy rooted in the node “entity”, and hypernymy-hyponymy (“is-a”) links as edges, since they are the most frequently encoded²⁰ and more likely to form a meaningful structure, resembling a taxonomy. Even with this simplification, the remaining nodes and edges do not define a simple tree, but define instead a *tangled tree*, i. e. a tree that contains a few nodes that can have multiple parents.

Par.	Nodes	Internal	Leaves	Types	Inst.
0	1	1	0	1	0
1	79,901	16,680	63,221	72,962	6,939
2	2,134	463	1,671	1,388	746
3	63	12	51	30	33
4	12	0	12	3	9
5	3	1	2	1	2
6	1	0	1	0	1
Total	82,115	17,157	64,958	74,385	7,730

Table 1: Count of synset nodes, grouped by number of parents.

We decided to consider the root synset 100001740 (entity, “That which is perceived or known or inferred to have its own distinct existence (living or nonliving)”) as the map center, and to place the remaining nodes around it, using a modified radial tree layout. For each synset, we computed the shortest path to reach the root

²⁰<http://wordnet.princeton.edu/>

node through hyponym links²¹ as a measure of tree depth. Then, we grouped the nodes in rings by depth value. Unlike a standard tree layout, we decided to place the nodes nearer to each other, by packing the nodes of a ring in more than one orbit. By using a traditional radial layout, in which all the nodes of a ring are placed in a single orbit, rings would have to be drawn very far from each other, because of the fast growth in the number of nodes as the ring index increases.

Ring	Nodes	Internal	Leaves	Types	Inst.
0	1	1	0	1	0
1	3	3	0	3	0
2	22	12	10	22	0
3	228	146	82	227	1
4	2,020	891	1,129	2,011	9
5	6,249	2,033	4,216	5,641	608
6	12,267	3,171	9,096	10,555	1,712
7	18,936	3,276	15,660	16,892	2,044
8	14,155	2,666	11,489	13,028	1,127
9	11,042	1,955	9,087	9,286	1,756
10	7,207	1,251	5,956	6,867	340
11	4,267	738	3,529	4,204	63
12	2,505	451	2,054	2,450	55
13	1,383	261	1,122	1,381	2
14	846	135	711	845	1
15	449	99	350	448	1
16	341	57	284	341	0
17	164	11	153	153	11
18	30	0	30	30	0
Total	82,115	17,157	64,958	74,385	7,730

Table 2: Count of synset nodes, grouped by ring. The ring index corresponds to the minimum depth of the node (length of the shortest path to the root node).

WordNet synsets are divided in two categories: types and instances. We identified and marked 7,730 instance noun synsets (target of instance hyponym links²²) to differentiate them visually in the map. We also marked the leaf nodes, 64,958 out of the total of 82,115 noun synsets (79%).

3.2 Implementation details

Displaying a graph with more than 80,000 nodes is a daunting task, even for state-of-the-art web technologies. Moreover, a visualization of a graph that big would probably be unsuitable for giving the user an overview of it. We decided to compute offline the fixed position of each synset node representation within the space of the graphic, and to

²¹Both regular and instance hyponym links, though in the visualization they are differentiated.

²²Type nodes are always destinations for regular hyponymy edges only, and instance nodes are destinations for instance hyponymy edges. This is true except for 5 specific nodes that are probably errors in the database.

store the resulting data in a geospatial database²³. This kind of databases provide spatial indices, to achieve good performance when issuing spatial queries. For example, a spatial query could request all the countries that overlap a specified polygon in a map. We implemented a REST-style (Fielding, 2000) web API that makes use of this technique to retrieve a JSON²⁴ representation of the nodes that overlap the rectangular region defined by the client viewport. This approach ensures that only the visible nodes are actually loaded from the server and displayed by the user agent, while giving to the user the illusion of having the whole graph loaded when panning.

When the user decreases the zoom, the viewport increases the covered region, and consequently the amount of information to load and show. As previously described, we followed a common approach used for Zooming User Interfaces, and reduced this amount of data by implementing a concept of Level Of Detail: each graphic element stored in the database is given a minimum zoom level for it to appear. The spatial database receives the current zoom setting of the client along with the position and size of the viewport, and discards graphic symbols that are considered too fine-grained to appear at that particular zoom level.

The JSON response returned by the web API is then processed client-side and transformed into an interactive vector graphic by a cross-browser²⁵ JavaScript library²⁶ that leverages HTML5 canvas²⁷ and Scalable Vector Graphics (SVG)²⁸ technologies.

The hypertextual dictionary view of WordNet is implemented by using similar techniques, applied to the single scroll dimension instead of the 2D space of the map.

The selection of words and synsets is implemented by following a common web application practice that requires to update the page hash URI whenever a meaningful change in application state is detected. In our work, this makes it possible to use the browser's history, navigation buttons and bookmarks to browse words and synsets, each uniquely identified by an URI. This enables the

user to perform the navigation task effectively, by leveraging proven user interaction paradigms and familiar interfaces. Each item presented by WordNet Atlas is thus uniquely identified by a dereferenceable URI²⁹ that can be shared or simply written in a document to unambiguously refer to the item.

4 Conclusions

We presented WordNet Atlas, a web application capable of giving WordNet users an interactive visualization of the resource. We believe that its capability to scale from a meaningful overview to the details of WordNet structure makes it ideal for presentation or teaching purposes, or for the promotion of WordNet to a broader audience. We also believe that the effort to investigate a graphical way to organize such a big graph could assist in the inspection and improvement of the underlying database, for example by exposing anomalies in the structure caused by human errors.

4.1 Future work

In the next step of our work, we will search for confirmation of the usefulness of the presented visualization technique. We are already setting up a round of user evaluation tests, with comparisons with interactive representations of WordNet found in literature.

In our future research, we hope to improve both the quality of the interaction and the amount of information the application is capable of conveying, while keeping its design polished. In respect to that, the visualization layout can probably be improved by taking into account more information, such as the node height (length of the path to a leaf) or the distinction between type and instance nodes. We plan to provide a visualization for the other three WordNet subnets, verbs, adjectives and adverbs, linked to the current one.

It could be interesting to use the same system behind WordNet Atlas to display wordnets in other languages, and to compare and interlink the resulting representations.

The described methodology could also prove to be useful to represent other big graphs, such as social graphs or semantic web graphs, and even data sets with a non-graph structure.

²³<http://postgis.refractions.net>

²⁴<http://www.json.org>

²⁵WordNet Atlas is reported to work in Mozilla Firefox 4+, Internet Explorer 8+ and Chrome 14+.

²⁶<http://raphaeljs.com>

²⁷<http://www.w3.org/TR/html5/the-canvas-element.html>

²⁸<http://www.w3.org/Graphics/SVG/>

²⁹This is a best practice in web development, commonly used in the semantic web initiative (<http://semanticweb.org>).

References

- James Abello, Jeffrey Korn, and Matthias Kreuseler. 2002. Navigating giga-graphs. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '02, pages 290–299, New York, NY, USA. ACM.
- David Auber. 2003. Tulip: A huge graph visualization framework. *Graph Drawing Softwares*, pages 105–4126.
- Christopher Collins. 2007. Wordnet explorer: Applying visualization principles to lexical semantics.
- Weiwei Cui. 2011. A survey on graph visualization.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.
- Roy Thomas Fielding. 2000. Architectural styles and the design of network-based software architectures.
- George W. Furnas. 1986. Generalized fisheye views. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 16–23.
- Sohaib Ghani and Niklas Elmquist. 2011. Improving revisitation in graphs through static spatial features. In *Proceedings of Graphics Interface 2011*, pages 175–182.
- Jaap Kamps. 2002. Visualizing wordnet structure. In *Proc. of the 1st International Conference on Global WordNet*, pages 182–186.
- George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38:39–41, November.
- Tamara Munzner. 2000. Interactive visualization of large graphs and networks.
- Ben Shneiderman. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343.
- Colin Ware. 2004. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Jiajie Zhang. 1991. The interaction of internal and external representations in a problem solving task. In *Proceedings of the Thirteenth Annual Conference of Cognitive Science Society*, pages 88–91.

Verbs in Sanskrit Wordnet

Tanuja Ajotikar
 Department of HSS
 IIT, Bombay
 [gtanu30@gmail.com]

Malhar Kulkarni
 Department of HSS
 IIT, Bombay
 [malhar-ku@gmail.com]

Pushpak Bhattacharya
 Department of CSE
 IIT, Bombay
 [pb@cse.iitb.ac.in]

Abstract

How far the language divergence affects the development of a Wordnet following the expansion approach, especially if languages are historically related and the Wordnet of the older language is created on the basis of the modern language, is the topic of discussion in this paper. The verbs in Hindi and Sanskrit differ from each other in nature which affects the process of developing the Sanskrit Wordnet (SWN) from Hindi Wordnet (HWN). We present here those differences and give strategies to create SWN. We seek to examine if the expansion approach forces to bring up issues rarely paid serious attention to so far.

1 Introduction

Sanskrit Wordnet (SWN) is being developed at IIT Bombay following the expansion approach, with Hindi Wordnet (HWN¹) as the source for wordnet. Hindi is a Neo Indo Aryan (NIA) language whereas Sanskrit has three different stages namely Old Indo Aryan (OIA), Middle Indo Aryan (MIA) and Modern Sanskrit. It would generally be unacceptable that the lexical database for a historically older language is developed on the basis of the database of the historically younger language as mentioned above. We show in this paper how in this scenario creation of SWN could prove useful linguistically.

1.1 Current status of the work

Category-wise break-up of the synsets in SWN is as follows-

Noun	7033
Adjective	2306
Adverb	170
Verb	735

Table.1 Account of 10244 synsets completed till 31st.8. 2011 with unique word count 28930 and ratio 3.2 per synset

Developing SWN with the expansion approach gives us an opportunity to study the linguistic peculiarities of Sanskrit from a fresh perspective. In this paper our focus will be on the divergence of verbs in Sanskrit and Hindi and its implication for constructing verbal synsets in SWN.

2 Strategies of creating verbal synsets in SWN

Below we discuss the strategies with regard to various issues.

2.1 Lexicalization vs. Suffixation

Hindi uses lexicalized units to denote tense, mood, and aspect. This kind of information is fused into one unit in Sanskrit and it is expressed by a suffix. The unit of a verb that bears such information is called as an ‘auxiliary’. Sanskrit has very few ‘auxiliaries’. Hence the synset which expresses such meaning cannot be mapped in SWN as the conveyor of such meaning *i.e.* suffix is not a part of lexicon. The example is given below-

1. (synset id. 10809) चुकना (chukanā) (Meaning- ‘to finish’)- किसी कार्य का बाकी न रहना (kisi kārya kā bākī na rahanā) (Meaning-) "क्या आप खाना खा चुके" (kyā āpa khānā khā chuke) (Meaning- Have you finished your meal?)

¹ <http://cfilt.iitb.ac.in>

चुकना (chukanā) is a modal auxiliary in Hindi. This particular sense is expressed through a suffix (लुङ्- lu~n or तवत्- tavat) in Sanskrit. Hence we cannot link this synset.

2. (synset id.10149) संभावना होना, सकना (saṁbhāvanā honā, sakānā) (Both of these verbs mean ‘to expect’ or ‘to anticipate.’) कोई कार्य हो सकना (koī kārya ho sakānā) (Meaning-regard something as probable.) आज बारिश होने की संभावना है। (ājā bāriśha hone kī saṁbhāvanā hai) (Meaning- It is likely to rain today.)

सकना (sakanā) is also a modal auxiliary in Hindi which is expressed through a suffix (लिङ्) in Sanskrit. Hence we cannot lexicalize above mentioned concepts in Sanskrit and link the synset of Hindi in HWN.

2.2 Light Verbs

Hindi has light verbs but they are absent in Sanskrit. Butt (2003:18) says that the decrease in the use of the pre-verbs has given rise to the use of the light verbs. Every language that uses the light verb has a closed set of such verbs. Hindi has 11² such verbs. These verbs express meanings like completion, suddenness, benefaction etc.

2.2.1 Light verbs and pre-verbs

The term pre-verb is applied to a morpheme which is attached to a verb. When a verb occurs with a pre-verb, its meaning sometimes changes. But pre-verb does not change its category.

Sometimes meanings like completion as in चला जाना (calā jānā) would be conveyed through a pre-verb in Sanskrit like निर् (nir) as in निर्गतः (nirgataḥ) (he went).

2.2.2 Reverse Example of pre-verb expressing the meaning of a light verb

Here is the reverse example where Sanskrit does not express the meaning of the light verb by any of the pre-verbs³.

<p>7122 एक चीज आदि के बदले में दूसरी चीज आदि लेना या देना (ek cīja ādi ke badale mein dūsari cīja ādi lenā yā denā) (Meaning- exchange or replace with another)</p> <p>रमा ने अपना पुराना फ्रीज बदल दिया (ramā ne apanā purānā freez badala diyā) (Meaning- Ramā exchanged her fridge.)</p> <p>बदलना, बदल देना (badalānā badala denā) (All the verbs mean ‘to exchange’)</p>	<p>एक वस्तु दत्त्वा अन्यस्य वस्तुनः ग्रहणानुकूलः व्यापारः। (ekam vastu datvā anyasya grahaṇānukūlaḥ vyāpāraḥ) (Meaning- exchange or replace with another)</p> <p>“रमा स्वस्य शीतकपाटिकां प्रत्ययच्छत्। (ramā svasya śītakapātiikām pratyayacchat) /तिलेभ्यः प्रतियच्छति माषाण्। (tilibhyah pratiyacchati māṣāṇ) (Meaning- Ramā exchanged her fridge.)</p> <p>प्रतिदा (pratidā), विमे (vime), निमे (nime), मे(me), प्रतिपण् (pratipan), व्यतिह (vyatihi), व्यवह (vyavah), परिवृत् (parivṛt) (All the verbs mean ‘to exchange’)</p>
--	---

Table.2

a pre-verb expressing meaning of a light verb

The concept in 7122, ‘to exchange’, can be expressed in Hindi with the help of two verbs. One of these verbs in Hindi is simple verb and other one is V+V i.e. compound verb. The second member देना (denā) is a light verb. In Sanskrit the concept [(एक चीज आदि के बदले में दूसरी चीज आदि लेना या देना) (ek cīja ādi ke badale mein dūsari cīja ādi lenā yā denā)] (meaning- ex-

³ These are as follows- प्र (pra), परा (parā), अप (apa), सम् (sam), अनु (anu), अव (ava), निस् (nis), निर् (nir), दुस् (dus), दुर् (dur), वि (vi), आङ् (ān), नि (ni), अधि (adhi), अपि (api), अति (ati), सु (su), उद् (ud), अभि (abhi), प्रति (prati), परि (pa-ri), उप³ (upa) (Total 22).

² Poornima (2008:7) There are 11 light verbs listed in the paper. बैठना (baiṭhanā), डालना (dālanā), देना (denā), लेना (lenā), मारना(māranā), निकालना (nikālanā), आना (ānā), जाना (jānā), पड़ना (paḍnā), निकलना (nikalānā), उठना (uthānā). One verb is repeated as they are grouped according to transitive and intransitive nature.

change or replace with another) is expressed by adding a pre-verb to a verb which means exactly opposite e.g. प्रतिदा (pratidā). The root दा (dā) means ‘to give’ but when प्रति (prati) is added to a verb then its meaning changes to ‘to exchange’. Same can be said about पण् (paṇ) which means ‘to buy’ and when प्रति (prati) is added to the verb then its meaning changes to ‘to exchange’. व्यतिहृ (vyatihṛ), व्यवहृ (vyavahṛ) also mean ‘to exchange’ but the verb occurs here with pre-verbs like वि+अति (vi+ati) and वि+अव (vi+ava). The meaning of the root हृ (hṛ) is ‘to take’ but it changes when these particular pre-verbs are added to the root. Sometimes the root expresses the same meaning with the pre-verbs too, as मे (me) itself means ‘to exchange’ but विमे (vime), निमे (nime) also mean the same.

All the members of the synset express the meaning of the root बदलना (badalanā) only. But no member of the synset in Sanskrit expresses the sense of देना (denā).

2.3 Conjunct verbs

The noun or adjective or adverb and a simple verb together form a conjunct verb.

2.3.1 Conjunct Verb with a तत्सम word in Hindi and Simple Verb in Sanskrit

The conjunct verbs with the combination of- a तत्सम⁴ word + simple verb, in HWN will be expressed by a single verb unit in Sanskrit, e.g.

इच्छा करना (icchā karanā), कामना करना (kāmanā karanā), इच्छा रखना (icchā rakhanā) (to desire)	इष् (iṣa), कम् (kam) (to desire)
सर्वनाश करना (sarvanāśa karanā) (to destroy)	नश् (naś), ध्वंस् (dhvamsa) (to destroy)
शान्त करना (śānta karanā) (to stop)	शम् (śam) (to stop)

Table.3 Conjunct Verb with a तत्सम word in Hindi and Simple Verb in Sanskrit

⁴ तत्सम word is a Sanskrit word used in modern languages.

These verbs in Hindi are conjunct verbs and the verbs in Sanskrit are the roots of the words in Hindi conjunct verbs, e.g.; इच्छा (icchā) is derived from the verbal root इष् (iṣa), कामना (kāmanā) is derived from कम् (kam) and so on.

2.3.2 Conjunct Verb in Hindi and Pre-verb+Verb in Sanskrit

7861 किसी का अनुसरण करना (kisi kā anusaraṇa karanā) (Meaning- to follow somebody)	पश्चात् धावनानुकूलः व्यापारः। (paścāt dhāvnānukūlh vyāpārah) (to follow somebody)
"पुलिस ने बहुत देर तक चोर का पीछा किया" (pulisa ne bahuta dera takha chor kā pīchā kiyā) (Meaning- The police followed the thief for long time.)	आरक्षकाः दीर्घकालं यावत् चौरम् अन्वसार्षः। (ārakṣakāḥ cauram anvasārshuh) (The police followed the thief for long time.)
पीछा करना, पछियाना, पीछे लगना (pīchā karanā pachiyānā pīche laganā) (All the verbs mean ‘to follow’)	अनुस् (anusṭ), अनुगम् (anugam), अनुधाव् (anudhāv), अनुव्रज् (anuvraj), अनुया (anuyā), अनुवृत् (anuvṛt), अनुक्रम् (anukram), अनुद्रु (anudru), समनुगम् (samanugam), समनुद्रु (samanudru) (All the verbs mean ‘to follow’)

Table.4 Conjunct Verb in Hindi and Pre-verb+Verb in Sanskrit

This example (id 7861) is in the sense of ‘to pursue or to follow’. In Hindi it is expressed through conjunct verb (पीछा करना ‘pīchā karanā’, पीछे लगना ‘pīche laganā’) and a simple verb (पछियाना ‘pachiyānā’). In Sanskrit, it is expressed through a pre-verb अनु (anu) and समनु (samanu) along with verbs which mean ‘to go’ and ‘to run’. Here two pre-verbs combine together. This is the unique feature of Sanskrit absolutely absent in Hindi.

2.4 Simple Verb in Hindi and Conjunct Verb in Sanskrit

Here is the example where Hindi expresses a particular concept through a simple verb. But Sanskrit, against its natural tendency, expresses a particular concept through a noun and simple verb combination.

5351 सूर्यं चंद्रं आदि का अस्तं होना (sūrya Chandra ādi kā asta honā) (Meaning- Setting of the sun or moon)	सूर्यस्य चन्द्रस्य वा अस्तं प्रति गमनानुकूलः व्यापारः। (Sūryasya chandrasya vā astarīn prati gamanānukūlāḥ vyāpāraḥ) (Meaning- setting of the sun or moon)
"सूर्यं पश्चिम में इबता है" (Sūrya paścima me dūbatā hai) (Meaning- The sun sets in the west.) इबना (dūbanā), ढलना (dhalanā), अस्तं होना (asta honā), अस्तंगत होना (astagata honā) (All the verbs mean 'to set')	"सूर्यः पश्चिमदिशि अस्तं गच्छति।" (Sūryah paścimadiśI astarīn gaccha- ti) (Meaning- The sun sets in the west.) अस्तं_गम् (astarīn_gam), अस्तं_व्रज् (astarīn_vraj), अस्तं_या (astarīn_yā), अस्तं_इ (astarīn_i), अस्ताचलम्_अवलंब् (astācalam_avalamb), अस्तशिखरम्_अवलंब् (as- taśikharam_avalamb), अस्ताचलम्_प्राप् (astācalam_prāp), अस्तशिखरम्_प्राप् (as- taśikharam_prāp), सागरे_मस्ज् (sāgare_masj) (All the verbs mean 'to set')

Table.5 Simple Verb in Hindi and Conjunct Verb in Sanskrit

This is an example of such concept which is not expressed through a single verb in Sanskrit. The synset in Hindi is 'इबना' (dūbanā) means 'to set'. This concept is expressed through a single verb, ढलना (dhalanā) as well as इबना' (dūbanā), in Hindi. Other members of the synset are 'conjunct verbs'. The first member is 'तत्सम' (tatsama) word in it. There should be a strategy to map this synset in SWN. The possible solution is, a phrase which means 'to set' in Sanskrit should be entered in the synset like HWN does in other synsets of verbs.

Should we call all the members (**अस्तं_गम**) (astarīn_gam) etc. of this synset in SWN 'conjunct verb'? The first word in 'astarīn gam' means 'setting' and second word means 'to go'. Both of

these words together mean 'to set'. 'Asta' is an independent lexeme in Sanskrit and 'gam' too. In this way one concept is expressed through 'noun+simple verb' combination. Hence it can be labeled as a 'conjunct verb'.

3. Emerging Issues

3.1 Role of pre-verbs

In the section 2.3.2 we have shown that Sanskrit expresses meanings (expressed by conjunct verbs in Hindi) through pre-verb and verb combination. Hence pre-verbs will play an important role as far as the lexicalization in Sanskrit is concerned. The pre-verbs are also important to study the relational semantics.

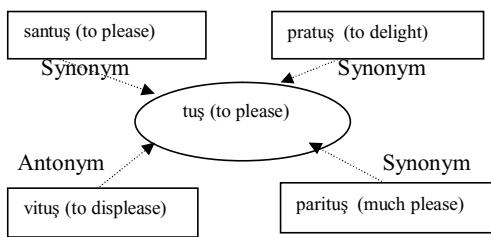


Fig.1 Semantics of pre-verb

The verb तुष् (tuṣ) means 'to satisfy' or 'to please'. When various pre-verbs like सम् (sam), प्र (pra), परि (pari) etc. are attached to it then meaning of the verb gets changed slightly, hence we may say that these newly formed words are synonyms of the original verbs⁵. But the other pre-verb i.e. वि expresses opposite meaning i.e. 'to be unhappy'.

The figure no.1 gives an over-view that such kind of study of the relations among the verbs and their combinations with the pre-verbs would reveal some interesting issues in the field of 'relational' semantics which forms a foundation to the Wordnet activity.

3.2 Nature of conjunct verbs in Sanskrit

Sen (reprint 1995) is an important contribution in this discussion. It provides (presented below) a list⁶ of verses in Mahābhārata⁷ and goes to show the usage of conjunct verbs.

⁵ This feature drastically increases the size of the synset.

⁶ Sen 1995:382

⁷ Mahābhārata is an epic composed by Vyāsa.

The list is given below.

1. राज्यं कारयितुम् Mahābhārata I.80.27
(rājyam kārayitum)
State to make somebody to do
(to rule a state)
2. पाण्डुभीर्वं चक्रे । Mahābhārata 1.90.73
(Pāṇḍurbhāvam cakre)
Pāṇḍu opinion do
(Pāṇḍu decided)।
3. गमनाय मतिं चक्रे । Mahābhārata 2.2.2
(gamanāya matim cakre)
To go **intellect to do**
((He) **decided** to go)।
4. राजसूये मनः कुरु । Mahābhārata 2.13.26
(rājasūye manah kuru)
rājasūya mind do
(Concentrate on rājasūya)
5. यज्ञायैव मनो दधे । Mahābhārata 2.13.4
(yajñāyāiva mano dadhe)
sacrifice only **mind keep**
(Concentrate only on sacrifice.)
6. वाचं ददानि । Mahābhārata 3.265.3
(vācam dadāni)
word give
((Let me) promise)

The members of each verb (in these cases) are a noun and a verb. The verbs like कृ (kṛ), दा (dā), धा (dhā) are used here. So we may say that there are ‘conjunct verbs’ found in the Sanskrit literature also. On the basis of this discussion we can say that we use conjunct verbs in SWN to map concepts in HWN.

We are also aware of the distinction between the conjunct verbs and the peri-phrases.

3.2.1 Peri-phrases- Most of the verbs in Sanskrit can be converted into ‘derived form of the verb + kr’ format (peri-phrase) e.g. गच्छति (gacchati)-
गमनं करोति⁸ (gamanan̄ karoti) (he goes), पठति
(paṭhati)- पठनं करोति (pathanam̄ karoti) (he studies), शेते (sete)- शयनं करोति (śayanam̄ karoti) (he

sleeps) etc. Because of their distinct nature from the conjunct verbs we do not make peri-phrases part of SWN

4. Conclusion and Future Work

Pre-verbs play an important role in Sanskrit and will be used effectively in SWN to express various Hindi language phenomenon like ‘light verbs’. It could be shown that Sanskrit has conjunct verbs. The study of the inchoative constructions in Sanskrit and conjunct verbs in Hindi is an interesting issue which we reserve for future discussions.

References

Bhattacharrya Pushpak, Chakrabarti, Debashri, and Sarma, Vaijayanti, 2007. Complex Predicates in Indian LangWordnets.

www.cse.iitb.ac.in/~pb/papers/Debasri-Complex-Predicates-26mar07.pdf

Butt, Miriam. 2003. The Light Verb Jungle. In Harvard Working Papers in Linguistics, ed. G. Aygen, C. Bowern, and C. Quinn. 1–49. Volume 9, Papers from the GSAS/Dudley House Workshop on Light Verbs. (<http://Ling-unikonstanz.de/pages/home/butt>)

Chakrabarti, Deashree. Mandelia, Hemang. Sarma, Vaijanthi. Bhattacharrya, Pushpak. 2008. Hindi Compound Verbs and their Automatic Extraction. Computational Linguistics, Manchester, UK
<http://www.ces.iitb.ac.in/~pb/pubs-yearwise.html>

Poornima, Shakthi, 2008. Reverse Complex Predicates in Hindi, A report submitted to Department of Linguistics SUNY University at Buffalo. (www-student.cse.buffalo.edu/~poornima/pdfs/qp.pdf)

Sen, Dr., Sukumar, 1995 (reprint). Syntactic Studies of Indo-Aryan Languages, Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign Languages.
(first publishing year not known)

Speijer, J. S. 1980 (reprint) (first edition 1885). Sanskrit Syntax, New Delhi: Motilal Banarasidas Publishers.

⁸ Speijer uses a term ‘Peri-phrase of a verb’ and notes that such kinds of constructions are frequent in the literature.

Using WordNet to Handle the OOV Problem in English to Bangla Machine Translation

Khan Md. Anwarus Salam
The University of Electro-
Communications
Chofu, Tokyo, Japan.
kmanwar@gmail.com

Yamada Setsuo
NTT Corporation
Tokyo, Japan.
yamada.setsuo@
lab.ntt.co.jp

Tetsuro Nishino
The University of Electro-
Communications
Chofu, Tokyo, Japan.
nishino@ice.uec.ac.jp

Abstract

Machine Translation (MT) for low-resource language has high probability of handling Out-Of-Vocabulary (OOV) words. Using WordNet we propose a method to handle the OOV problem in Example-Based Machine Translation (EBMT) for English to Bangla language. Proposed method first tries to find semantically related English words from WordNet for OOV. From these related words, we choose the semantically closest related word whose Bangla translation exists in English-Bangla dictionary. If no Bangla translation exists, the system uses IPA-based-transliteration. For proper nouns, the system uses Akkhor Bangla transliteration mechanism.

1 Introduction

Machine Translation (MT) for low-resource language has high probability of handling Out-Of-Vocabulary (OOV) words. According to “Distribution of languages on the Internet”¹, 56.4% web contents are in English. On the other hand, “Human Development Report 2009”² of United Nations Development Program (UNDP) reported the literacy rate of Bangladesh as 53.5%. That means, around half of the Bangla speaking people of Bangladesh are monolingual. To improve the information access to those Bangla speaking monolingual people, it is important to have good English to Bangla Machine Translation (MT) system. However, lack of parallel corpus makes the development of the MT system very challenging.

English has rich language resources like automated parser and tokenizer. WordNet is a large lexical database of English [4]. On the other

hand Bangla is a low-resource language due to the lack of language resources like Bangla WordNet and authorizedparallel corpus.

In this situation, to utilize the available language resources for English, we consider to use English as source language (SL) and Bangla as target language (TL). The Proposed method only uses SL analysis information and it does not require rich TL resources.

There were several attempts at building English-Bangla MT systems. The first available free MT system from Bangladesh was Akkhor Bangla Software³. The second available online MT system was apertium based Anubadok⁴. These systems used Rule-Based approach and did not handle OOV Words considering low-resource scenario. Most recently from June 2011, Google Translation⁵ started offering MT service for Bangla language, having issues in translating OOV Words.

In present, there are several approaches for Machine Translation, such as Rule-Based MT (RBMT), Statistical MT (SMT) and Example-Based MT (EBMT). RBMT require human made rules, which are very costly in terms of time and money, but still unable to translate general-domain texts. SMT and EBMT both are data driven approach. SMT works well for close language pairs like English and French. It requires huge parallel corpus, but currently huge English-Bangla parallel corpus is not available. EBMT is better choice for Bangla language, as it is less demanding on large parallel corpus. Moreover, EBMT system can translate in good quality when it has good example match. All these approaches has issues on translating OOV Words.

We considered EBMT approach by improving the translation quality using WordNet. For using

¹ <http://www.netz-tipp.de/languages.html>

² <http://hdr.undp.org/en/reports/global/hdr2009/>

³ <http://www.akkhorbangla.com>

⁴ anubadok.sourceforge.net

⁵ <http://translate.google.com/#en|bn>

WordNet in translation rules we used chunk-string templates (CSTs) (Salam et. al, 2011). CSTs consist of a chunk in the source language (English), a string in the target language (Bangla), and the word alignment information between them. CSTs are generated from the aligned parallel corpus and WordNet, by using English chunker. For clustering CSTs, we used <lexical filename> information for each words, provided by WordNet-Online⁶.

Using WordNet we propose a method to handle the OOV problem in Example-Based Machine Translation (EBMT) for English to Bangla language. To improve the translation quality, we propose a method for EBMT using WordNet, IPA-based and Akkhor transliteration. Proposed method first tries to find semantically related English words from WordNet for the OOV words. From these related words, we choose the semantically closest related word whose Bangla translation exists in English-Bangla dictionary. If no Bangla translation exists, the system uses International-Phonetic-Alphabet(IPA)-based-transliteration. For proper nouns, the system uses the transliteration mechanism provided by Akkhor Bangla Software. Based on the above methods, we built an English-to-Bangla EBMT. The evaluation of the initial system is described (Salam et al., 2011).

2 Background

For this research we used EBMT approach for low-resource language using chunk-string templates (Salam, 2011). The Figure 1 shows the EBMT architecture. The dotted rectangles identified the main contribution area of this research. During the translation process, at first, the input sentence is parsed into chunks using OpenNLP Chunker. The output of Source Language Analysis step is the English chunks. Then the chunks are matched with the example base using the Matching algorithm as described in section IV. This process provides the CSTs candidates from the example-base. It also marks the OOV Words. In OOV Word Translation step, using our proposed mechanism in section V, we try to find translation candidates for those OOV Words. Then in Generation process WordNet helps to translate determiners and prepositions correctly to improve MT performance [7]. Finally using the generation rules we output the target-language strings. Based on the above MT system

architecture, we built an English-to-Bangla MT system.

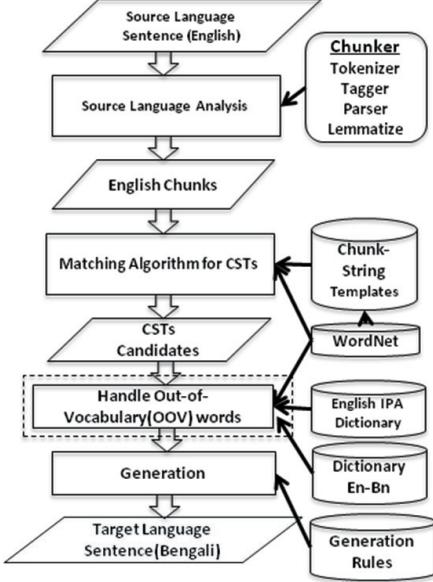


Figure 1: EBMT Architecture

In this research we used EBMT based on chunk-string templates (CSTs), which is especially useful for developing a MT system for high-resource to low-resource language. CSTs consist of a chunk in the source language (English), a string in the target language (Bangla), and the word alignment information between them. From the English-Bangla aligned parallel corpus CSTs are generated automatically.

English	Bangla	Align
Bangla is the native language of	বিশ্বব্যাপী বাংলা	11 1
1 2 3 4 5 6	হচ্ছে প্রায় ২৩০	2 7 8
around 230 million people worldwide	মিলিয়ন মানুষ —এর	9 10 6
7 8 9 10 11	মাহাত্মা	4

Table 1: Example word-aligned parallel corpus

Table 1 shows sample word-aligned parallel corpus. Here the alignment information contains English position number for each Bangla word. For example, the first Bangla word “বিশ্বব্যাপী” is aligned with the 11th word in the English sentence. That means “বিশ্বব্যাপী” is aligned with “worldwide”.

The example-base of our EBMT is stored as CSTs. We produced CSTs from the parallel corpus. Table 2 shows the initial CSTs for the parallel sentence given in Table1. In Table 2, c is a chunk in the source language (English), s is a string in the target language (Bangla), and t is the alignment information calculated from the original word alignment.

⁶ <http://wordnetweb.princeton.edu/perl/webwn>

CST#	English Chunk (C)	Bangla (S)	T
CST1	[NP Bangla/NNP]	বাংলা	1
CST2	[VP is/VBZ]	হচ্ছে	1
CST3	[NP the/DT native/JJ laguage/NN]	মাতৃভাষা	2
CST4	[PP of/IN]	-এর	1
CST5	[NP around/RB 230/CD million/CD people/NNS]	প্রায় ২৩০ মিলিয়ন মানুষ	1 2 3 4
CST6	[ADVP worlwide/RB]	বিশ্ববাসী	1

Table 2: Example of initial CSTs

In the next step CSTs are generalized by using WordNet to increase the EBMT coverage. To generalize we only consider nouns, proper nouns and cardinal number (NN, NNP, CD in OpenNLP tagset). For each proper nouns we search in WordNet. If available we replace that NNP with <lexical filename> returned from the WordNet. For example WordNet return <noun.communication> for “Bangla”. For cardinal number we simply CDs together to <noun.quantity>. We show example generalized CSTs produced using WordNet in Table 3.

CST#	English Chunk (C)	Generalized Chunk
CST1	[NP Bangla /NNP]	[NP<noun.communication>/NNP]
CST5	[NP around/RB 230/CD million/CD people/NNS]	[NP around/RB <noun.quantity> people/NNS]

Table 3: Combined- CSTs examples

Finally we get the CSTs database which has three tables: initial CSTs, generalized CSTs and Combined-CSTs. From the example word-aligned parallel sentence of Table 1, system generated 6 initial CSTs, 2 Generalized CSTs and 4 Combined-CSTs.

2.1 Matching Algorithm for CSTs

From the set of all CSTs we select the most suitable one, according to the following criteria:

1. The more CSTs matched, the better;
 2. Linguistically match give priority by following these ranks, higher level is better:
- Level 4: Exact match.
 - Level 3: <lexical filename> of WordNet and POS tags match
 - Level 2: <lexical filename> of WordNet match
 - Level 1: Only POS tags match
 - Level 0: No match found, all OOV words.

3 Handle Out-of-Vocabulary Problem

As in our assumption, the main users of this EBMT will be monolingual people; they cannot read or understand English words written in English alphabet. However, with related word translation using WordNet and Transliteration can give them some clues to understand the sentence meaning. As Bangla language accepts foreign

words, transliterating an English word into Bangla alphabet, makes that a Bangla foreign word. For example, in Bangla there exist many words, which speakers can identify as foreign words.

Figure 3 shows the OOV or Unknown Words translation process in a flow chart. Proposed system first tries to find semantically related English words from WordNet for the unknown word. From these related words, we choose the semantically closest related word whose Bangla translation exists in English-Bangla dictionary. If no Bangla translation exists, the system uses IPA-based-transliteration. For proper nouns, the system uses transliteration mechanism provided by Akkhor Bangla Software.

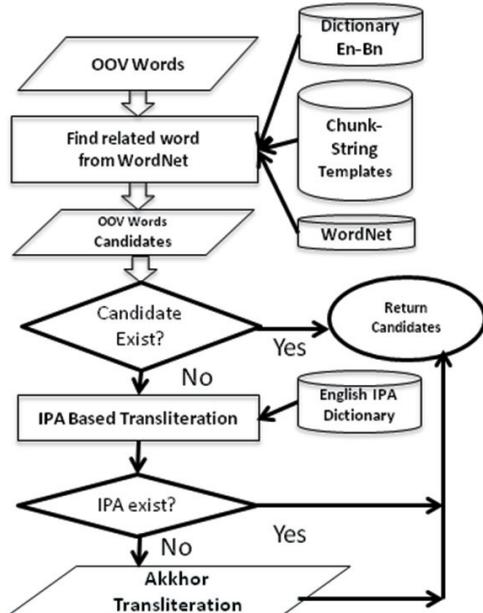


Figure 2: Steps of handling OOV words.

3.1 Find related words from WordNet

As small English-Bangla parallel corpus is available, there is high probability of translating OOV words. Therefore, it is important to have a good method for translating OOV words. When the word has no match in the CSTs, it tries to translate using English WordNet and bilingual dictionary for English-Bangla.

Input of this step is OOV or unknown words words. For example “dog” is a OOV in our system. Output of this process is the related OOV words translation.

3.1.1 Find Translation Using Synonyms

WordNet provide related word for nouns, proper nouns, verbs, adjectives and adverbs. The system

first finds the synonyms for the OOV word from WordNet. If any of the synonyms exist in English-Bangla dictionary, the system translate the OO word with that synonym meaning.

Polysemous words need word sense disambiguation techniques to choose the right candidate of the OOV word. For translating polysemous words we used special rules (Salam et. al. 2009).

3.1.2 Find Translation Using Hypernyms

For nouns and verbs WordNet provide hypernyms, which is defined as follows:

Y is a hypernym of X if every X is a (kind of) Y.

For example “canine” is a hypernym of noun “dog”, because every dog is a member of the larger category of canines. Verb example, “to perceive” is an hypernym of “to listen”. However, WordNet only provides hypernym(s) of a synset, not the hypernym tree itself. As hypernyms can express the meaning, we can translate the hypernym of the unknown word. To do that, until any hypernym’s Bangla translation found in the English-Bangla dictionary, we keep discovering upper level of hypernym’s. Because, nouns and verbs are organized into hierarchies, defined by hypernyms or is-a-relationships in WordNet. So, we considered lower level synset words are generally more suitable then the higher level synset words.

This process discovers the hypernym tree from WordNet in step by step. For example, from the hypernym tree of “dog” from WordNet, we only had the “animal” entry in our English-Bangla dictionary. Our system discovered the hypernym tree of “dog” from WordNet until “animal”. Following is the discovered hypernym tree:

dog, domestic dog, Canis familiaris
=> canine, canid
=> carnivore => placental, placental mammal
=> mammal => vertebrate, craniate => chordate
=> animal => ...

This process search in English-Bangla dictionary, for each of the entry of this hypernym tree. So at first we used the IPA representation of the English word from our dictionary, then using transliterating that into Bengali. Then system produce “a kind of X” - এক ধরনের X [ek dhoroner X].

For example of “dog” we only had the Bengali dictionary entry for “animal” form the whole hypernym tree. We translated “dog” as the translation of “dog, a kind of animal”, in Bangla

which is “ডগ, এক ধরনের পশু” [dog, ek dhoroner poshu].

Similarly, for adjectives we try to find “similar to” words from WordNet. And for Adverbs we try to find “root adjectives”.

Finally, this step returns OOV words candidates from WordNet which exist in English-Bangla dictionary.

3.2 Transliteration

When OOV word is not even found in WordNet, we use IPA-Based transliteration using the English IPA Dictionary (Salam, 2011). Output for this step is the Bangla word transliterated from the IPA of the English word. In this step, we use English-Bangla Transliteration map to transliterate the IPA into Bangla alphabet.

However, when unknown word is not even found in the English IPA dictionary, we use transliteration mechanism of Akkhor Bangla Software. For example, for the word “Muhammad” which is a popular Bangla name, Akkhor transliterated into “মুহাম্মদ” in Bangla.

4 Example

The implementation details and the evaluation of the EBMT-baseline is described (Salam et al., 2011). Table 6 shows sample translation examples produced by EBMT-Baseline and EBMT with OOV words solution. It also shows the translation quality in bracket (A,B,C,D: Perfect, Good, Medium, Poor).

#	English	EBMT Baseline	EBMT with OOV words
1.	Are you looking for an aardvark?	আপনি কি aardvark খুঁজছেন?(D)	আপনি কি আর্ডভার্ক, এক ধরনের পশু খুঁজছেন?(A)
2.	This dog is really cool.	Dog আসলেই দারলন (D)	ডগ, এক ধরনের পশু আসলেই দারলন (A)
3.	I am eating onigiri	আমি onigiri খাচ্ছি (D)	আমি অনিগিরি খাচ্ছি(A)
4.	His name is Rupok.	তার নাম Rupok. (D)	তার নাম রূপক(A)
5.	What is abstraction?	abstraction কি? (D)	এ্যাবস্ট্রিকশান কি? (B)

Table 6: Proposed system produced translations comparison with the baseline system

As “aardvark” and “dog” are OOV words, EBMT baseline produced poor translation for #1 and #2. Our proposed solution improved these two translations into perfect category. “onigiri” is a Japanese food name and “rupok” which is a person name are OOV words in #3 and #4. Our Akkhor transliteration solution improved these from poor to perfect translation. In the case of #5,

“abstraction” is an unknown word, which the system translated using the proposed IPA-based transliteration solution. As a result the translation improved to good quality. All these examples demonstrate the effectiveness of our proposed solution for translating OOV words.

5 Conclusion

We proposed to Use WordNet to handle the OOV problem in Example-Based Machine Translation (EBMT) for English to Bangla language. Proposed method first tries to find semantically related English words from WordNet for OOV. From these related words, we choose the semantically closest related word whose Bangla translation exists in English-Bangla dictionary. If no Bangla translation exists, the system uses IPA-based-transliteration. For proper nouns, the system uses Akhhor Bangla transliteration mechanism. Although current system works well for small parallel corpus. In future, we want to do the detailed analysis of the proposed method with larger test-set.

References

- Abney, Steven. 1991. Parsing by chunks. In Principle-Based Parsing, pages 257–278. Kluwer Academic Publishers.
- Diganta Saha, Sivaji Bandyopadhyay. 2006. A Semantics-based English-Bengali EBMT System for translating News Headlines. Proceedings of the MT Summit X, Second workshop on Example-Based Machine Translation Programme.
- Diganta Saha, Sudip Kumar Naskar, Sivaji Bandyopadhyay. 2005. A Semantics-based English-Bengali EBMT System for translating News Headlines, MT Xummit X.
- George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.
- Jae Dong Kim, Ralf D. Brown, Jaime G. Carbonell. 2010. Chunk-Based EBMT. EAMT, St Raphael, France.
- Khan Md. Anwarus Salam, Mumit Khan and Tetsuro Nishino. 2009. Example Based English-Bengali Machine Translation Using WordNet. TriSAI, Tokyo.
- Khan Md. Anwarus Salam, Yamada Setsuo and Tetsuro Nishino. 2010. English-Bengali Parallel Corpus: A Proposal. TriSAI, Beijing.
- Khan Md. Anwarus Salam, Setsuo Yamada and Tetsuro Nishino. 2011. Example-Based Machine Translation for Low-Resource Language Using Chunk-String Templates, 13th Machine Translation Summit, Xiamen, China.
- R. Gangadharaiyah, R. D. Brown, and J. G. Carbonell. Phrasal equivalence classes for generalized corpus-based machine translation. In Alexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, volume 6609 of Lecture Notes in Computer Science, pages 13–28. Springer Berlin / Heidelberg, 2011.
- Sajib Dasgupta, Abu Wasif and Sharmin Azam. 2004. An Optimal Way Towards Machine Translation from English to Bengali, Proceedings of the 7th International Conference on Computer and Information Technology (ICCIT).
- Sudip Kumar Naskar and Sivaji Bandyopadhyay. 2006b. Handling of Prepositions in English to Bengali Machine Translation. In the proceedings of Third ACL-SIGSEM Workshop on Prepositions, EACL 2006. Trento, Italy.
- Zhanyi Liu, Haifeng Wang And Hua Wu. 2006. Example-Based Machine Translation Based on Tree-string Correspondence and Statistical Generation. Machine Translation, 20(1): 25-41

A new hierarchy of ArchiWordNet (AWN): building parts implementation with image

Anna Rita Bertorello
Politecnico di Torino - DICAS
Torino, Italy
annarita.bertorello@polito.it

Abstract

Our paper presents a new development of a new hierarchy of ArchiWordNet (AWN), a bilingual thesaurus for architecture and building, concerning building elements. This thesaurus is being created according to the WordNet model and fully integrated in it. We describe the application of a regional law during the validation of the cataloguing cards of Guarini Patrimonio Culturale database; the compilation of the lexical hierarchy <building element, construction entity part> and its implementation with images.

1 Introduction

The ArchiWordNet (AWN) project has been developed by Politecnico di Torino – DICAS and Bruno Kessler Foundation (FBK). This collaboration allows to join competence in architecture (Politecnico) and computational linguistics (FBK) with the aim to create a thesaurus for architecture and building according to the WordNet model and fully integrated with WordNet itself (Bentivogli et al., 2004).

The first facets that we developed contain over 3.800 synsets distributed in <buildings and buildings complexes>, <building element, construction entity part>, and <material>.

We tested the adaptation of AWN terminology to indexing purposes in two cases: some architectural fonds at the Laboratory of History and Cultural Heritage of DICAS, and SIS¹ (Still Image Server) archive (Cavaglià, 2001); (Bocco et al., 2009).

The project has been further developed with the compilation of the new hierarchy <building

element, construction entity part>, which is being enriched with images.

2 The development of ArchiWordNet (AWN) applied to LR 35/1995

The development of AWN was supported by Regione Piemonte who commissioned us to validate eight fields of the cataloguing cards (*Guarini Patrimonio Culturale* database). Regione Piemonte also needed a thesaurus to organize the terminology of the regional architectural heritage census (Regional Law no. 35/1995), founded on the standards established by the Italian Central Institute for Cataloguing and Documentation (ICCD).

At the beginning, there was a phase of corpus analysis required to organize the lexical/morphological heterogeneity of the 30.359 compiled cards. These cards were commissioned by 672 Piedmontese municipalities and were compiled by hundreds of different professionals.

The analysis of *Guarini Patrimonio Culturale* records, allowed us to extract new terms to develop a new lexical hierarchy of AWN regarding <building element, construction entity part> and to obtain a feedback on the <buildings and buildings complexes> hierarchy, that had already been realized.

After the automatic extraction of the lemmas, we identified and organized manually possible repetitions and synonymies among the 24.267 lemmas allocated in eight fields.

During the validation operations we verified the terms consulting discipline-specific literature; we attributed the terms to correct synsets; we created a list of approved terms; we cross-checked AWN for possible integrations or corrections; we defined automatic procedures for

¹ SIS is an online database for architectural photographs created for didactic purposes at the Politecnico di Torino. This database can be reached by Politecnico users only.

the correction of inappropriate terms and we updated the *Guarini* user's guide.

During this work we met some problems concerning the use of an inappropriate lexicon, repetitions not recognised by the system ("door" ≠ "Door"), terms containing mathematical symbols ("fountain/wash-house"), misspelled words ("bowwindows") or terms not fitting the field. We also examined the photographs attached to many of the cards to understand the description of the object.

3 The development of the hierarchy <building element, construction entity part>

After the validation operations, the extracted terms were added to other words coming from norms, oral sources and discipline-specific literature, to enrich the lexicon of building parts (including, among others, Leva Pisto et al., 1993; Galliani, 2001; Maruffi, 1941; Pevsner, 1966; Hubner, 2003). Terms were not collected using automatic extraction tools both because when we started working on ArchiWordNet in 1999 we were not aware of the existence of such tools, and because we believe that, in our case, this method doesn't guarantee more satisfactory results, both in pertinence and quantity, in comparison to the manual method (Verlardi et al., 2005, Sagri et al., 2004, Stamou et al., 2002). In fact, the sources we have been using are high quality, domain-specific, and very synthetic.

To define the structure of the hierarchy we consulted some available classifications like the ISO 12006-2 norm (ISO, 2007), the UNI norm 8290-1 (UNI, 1981), the *Art & Architecture Thesaurus* (AAT) (Petersen, 1990), and other sources such as (Ray-Jones, 1991). During the population of this classification with synsets and relations we adopted the criterion of the functional role of each building part, as defined by ISO Technical Report no. 14.177 (ISO, 1994), and we had some feedback on the framework to modify it.

The hierarchy <building element, construction entity part> was structured in hypernymy/hyponymy and meronymy relations to describe the building. In fact, for our discipline it is important both that "A" is a type of B (ISA) and especially that it belongs to (an) other element(s) (IS PART OF).

For the construction of synonymy relationships, we started from the terms extracted from

LR 35/1995 cards which were related with other consulted sources to arrange synonymous terms depending on the frequency of use. To explain this, LR 35/1995 compilers used frequently some terms like "angular" than other synonymous terms founded in architectural literature like "quoins" or "corner".

All the terms collected were divided into groups ("structure", "enclosure", "partition", "architectural ornament", "service, apparatus", "instrumentality", and "building part"), that became direct hyponyms of <building element, construction entity part> as be seen in figure 1².

We also created two artificial nodes (<building part (by form)>, <building part (by position)>) to make the organization of <building part> hypernym easier.

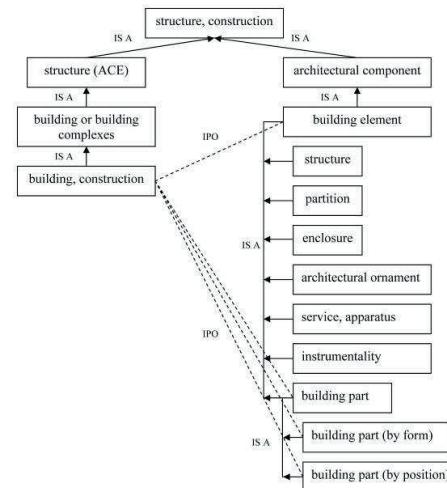


Figure 1. Structure of <building element, construction entity part>

To understand the importance of semantic relations for our discipline, we present an excerpt from <staircase, stairway, stairs> hierarchy, showing both hypernymy/hyponymy and meronymy relations. In figure 2, is represented the hierarchy of a tread in the following mode: a <tread, surface of tread> IS PART OF (IPO) <step, gradine, gradin, stepping>; <step, gradine, gradin, stepping> IS PART OF <staircase, stairway, stairs>; <staircase, stairway, stairs>

² Continuous lines indicate hypernymy/hyponymy relations (IS A) while broken lines meronymy relations (IS PART OF).

IS A <sloped partition>; <sloped partition> IS A <partition>; <partition> IS A <building element, construction entity part>; etc.

During the population of this hierarchy, we verified that a synset can be linked to other synsets by meronymy relations. In fact, <handrail, rail, hand railing>, IS PART OF <staircase, stairway, stairs>, IS PART OF <interior protection element, indoor protection element>, and IS PART OF <external partition element>.

We arranged 771 synsets and 1150 lemmas as hyponyms of <structure>, <partition> and <enclosure>. Our work is not finished and we are still populating the hyponyms <architec-

tural ornament>, <service, apparatus>, <instrumentality> and <building part>. Wherever possible, during this work we have preferred using synsets existent in WordNet and tag them as pertinent to the Architecture and Building Domain.

During the compilation of the <building element, construction entity part> hierarchy, we often hand-sketched some graphics to represent the objects and/or to make the gloss understandable. This is very common because through this instrument it is possible to mentally associate the object and the words that conceptualize it.

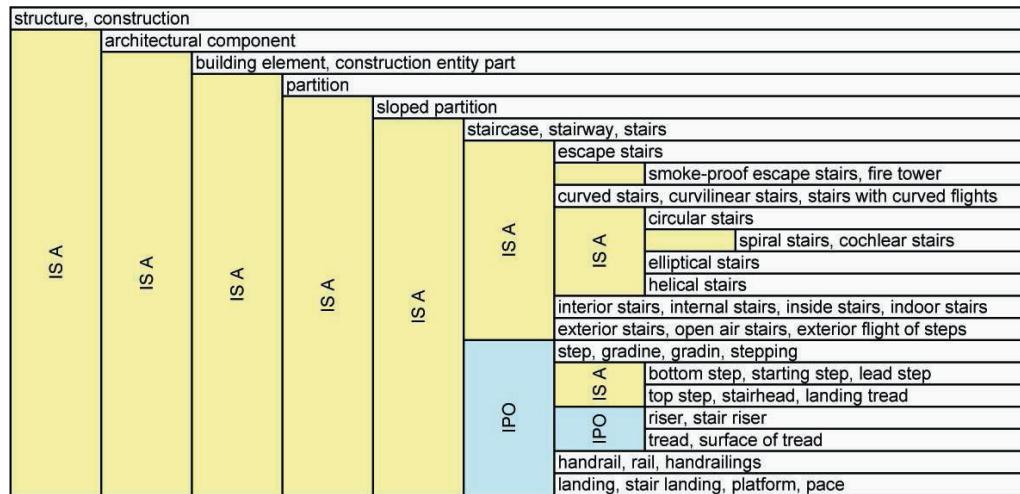


Figure 2. Structure of <staircase, stairway, stairs> relations

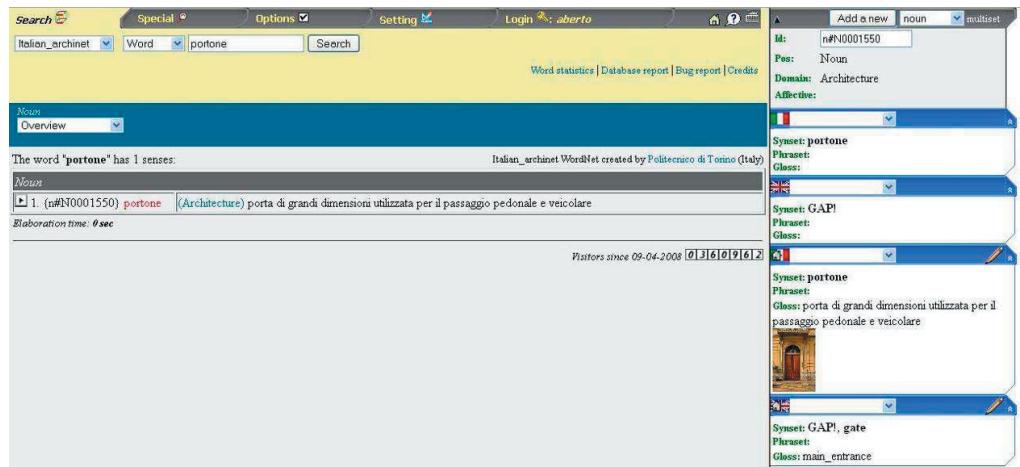


Figure 3. MultiWordNet interface

Practising this habit, with the help of FBK, we have been able to modify the interface of MultiWordNet to include photographs and captions with the purpose to realize an illustrated WordNet.

Now, we can upload images associated to single synsets. The system recognizes every single image because it adds a prefix - the number of the synset - before the name of the image file. For instance, if we have "portone.jpg" and we want to add this file to synset {n#N0001550} *main door*, the system will save the image with this name: "n_N0001550-portone.jpg". In this way, it is possible to link an image to different synsets, giving it different names.

In MultiWordNet interface it is possible to visualize the synset number, the synset, the glosses and the image(s). In figure 3, the interface of MultiWordNet is shown.

As a result of SIS cataloguing work and our habit to comment architectural photos with notes, we decided to associate a caption to each photo. In figure 4, it is shown that the caption can become apparent by the positioning of the mouse over the image.

It is also possible to enlarge the photo opening it in a new window.

These late developments allow to implement ArchiWordNet and make it easier to grasp the meaning of a synset (set of words) through the use of architectural images.

The <building element, construction entity part> hierarchy has been developed on the basis of the Italian lexicon, on the assumption that in a technical domain as that of architecture and building the language is relatively standardised and that semiotic gaps met in translation are few. The next step – still in progress – has been the completion of synsets with English terms and definitions.



Figure 4. The visualization of an image's caption (main door with impending fan-window)

4 Conclusions

In this article we have introduced our experience with the population of the <building element, construction entity part> hierarchy and the first attempt to realize an illustrated ArchiWordNet.

This instrument allows us to organize the concepts according to a classification even with image implementation. We demonstrated that it is possible to use this instrument in an application to a regional law.

In the future, we will continue to populate the remaining parts of AWN and to implement images.

Acknowledgements

I wish to thank Prof. Andrea Bocco and Prof. Gianfranco Cavaglià for their support and collaboration during the development of ArchiWordNet.

References

- Bentivogli L., Bocco A., Pianta E. 2004. *ArchiWordNet: Integrating WordNet with Domain-Specific Knowledge*, Proceedings of the 2nd Global WordNet Conference, Brno: Masaryk University
- Bocco A., Bodrato E., Perin A. 2009. *ArchiWordNet, a bilingual thesaurus for architecture and building: compiling and application to hybrid archives*, Proceedings of NAI Conference, Rotterdam
- Ray-Jones A., Legg D. 1991. *CI/SfB Construction Indexing Manual 1976*, London : RIBA Publications = Vetriani G., Marolda M.C. 1983, *Piano di classificazione PC/SfB*, Milano: ITEC editrice
- ISO 1994. *Technical Report 14177:1994. Classification of Information in the Construction Industry*, Geneva: International Organization for Standardization
- ISO 2007. *Standard 12006-3:2007. Building construction – Organization of information about construction works – Part 3: Framework for object-oriented information*, Geneva: International Organization for Standardization
- UNI 8290-1. 1981. *Edilizia residenziale. Sistema tecnologico. Classificazione e terminologia*
- Huber R., Rieth R. 1985-2003. *Glossarium Artis*, Munchen: K. G. Saur
- Pevsner N., Fleming J., Honour H. 1966. *The Penguin Dictionary of Architecture*, Harmonds-

- worth: Penguin = 1981. *Dizionario di architettura*, Torino: Einaudi
- Leva Pisto M., Molino M., Piovesana M.M. 1993. *Il Nomenclatore di Architettura*, Torino: Rosenberg & Sellier
- Petersen T. 1990. *Art and Architecture Thesaurus*, New York: Oxford University Press
- Galliani G.V. 2001. *Dizionario degli elementi costruttivi*, Torino: UTET
- Maruffi P. G. 1941. *Elementi costruttivi*, Padova: Zannoni Editore
- Cavaglià G. 2001. *L'analisi fotografica e la comprensione del costruito. Dalle patologie edilizie al progetto tecnologico*, Torino: Celid
- Velardi P., Navigli R., Cucchiarelli A., Neri F. 2005. *Evaluation of OntoLearn, a Methodology for Automatic Learning of Domain Ontologies*, Series information for Frontiers in Artificial Intelligence and Application, Amsterdam: IOS Press
- Stamou S., Ntoula A., Hoppenbrouwers J., Saiz-Noeda M., Christodoulakis D. 2002. *EUROTERM: Extending the EuroWordNet with Domain-Specific Terminology Using an Expand Model Approach*, Proceedings of the 1st Global WordNet Conference, Mysore: Central Institute of Indian Languages
- Sagri M.T., Tiscornia D., Bertagna F. 2004. *Jur-WordNet*, Proceedings of the 2nd Global WordNet Conference, Brno: Masaryk University

Introduction to Gujarati WordNet

Brijesh S Bhatt^{*†} C. K. Bhensdadia[‡] Pushpak Bhattacharyya^{*}
Dinesh Chauhan[‡] Kirit Patel[‡]

^{*}Center for Indian Language Technology
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay, India

[‡]Department of Computer Engineering
Faculty of Technology,
Dharmsinh Desai University, Nadiad, India

Abstract

Gujarati is one of the 22 official languages of India. Gujarati WordNet is being built using expansion approach with Hindi as the source language. This paper describes experiences of building Gujarati WordNet. Paper discusses basic features of Gujarati language and evaluates suitability of Hindi language for expansion approach. Various issues related to synset linking using expansion approach and challenges related to language specific concepts are also discussed.

1 Introduction

Wordnets have emerged as a very useful resource for computational linguistics and many natural language processing applications. Since the development of Princeton WordNet (Fellbaum C., 1998), WordNets are being built in many other languages. Hindi Wordnet(Narayan D. et al., 2002) is the first WordNet built for an Indian language. Based on Hindi WordNet, WordNets for 17 different Indian languages are being built. One such effort is Gujarati WordNet.

The paper is organized as follows, section 2 provides brief introduction to Gujarati language, section 3 describes influence of other languages on Gujarati. Synset development approach and synset categories are discussed in Section 4 and 5, respectively. Section 6

gives the current status of Gujarati WordNet. Issues related to synset linking are discussed in section 7. Section 8 describes a plan to validate synsets developed.

2 Gujarati Language

Gujarati, a native language of Indian state of Gujarat, is a member of Indo-Aryan family of languages. There are over 50 million speakers of Gujarati language.

Initially, the writing system of Gujarati was restricted to business writing , while the literature was in Devanāgarī script. The poetry form of the language is much older, enriched by poetry of poets like Narsinh Mehta. Gujarati prose writing and journalism started in 19th century. Protest writing against colonialism led to a string of powerful essays leading to the foundation of modern Gujarati literature.

Some features of Gujarati language are as follows:

2.1 Writing system

Gujarati script is a variant of Devanāgarī script, differentiated by the loss of the characteristic horizontal line running above the letters and by a small number of modifications in the remaining characters.

e.g. ‘kamal’ is written as,
Hindi: कमल, Gujarati: કમળ

2.2 Vocabulary

As Gujarati is an Indo-Aryan language descended from Sanskrit, its vocabulary con-

tains four general categories of words: *tat-sama*, *tadbhava*, *deshi* and *videshi* words.

- *tatsama*: Set of words accepted from Sanskrit language.
- *tadbhava*: Set of words from Sanskrit language adopted with a change in the phonological form.
- *deshi*: Words which are specific to Gujarati Language.
- *videshi*: Words which are accepted from different languages, like Persian, English, Portuguese etc.

It is also noteworthy that in some cases *tat-sama* and *tadbhava* words for a Sanskrit word co-exist with same or different meanings. For e.g. (1) ધર્મ (Dharma) and ધરમ (Dharam) both means same, 'Religion'. While, (2) કર્મ (karma) means Work, with religious connotation and કરમ (karam) means Work in general sense.

2.3 Grammar

Gujarati follows Subject-Object-Verb word order. There are three genders and two numbers. There is no article. Some significant features are as follows:

2.3.1 Gender

Gujarati distinguishes between three genders, masculine, feminine and neutral. For e.g.

છોકરો (chhokaro , Boy)

છોકરી (chhokarI , Girl)

છોકડુ (chhokarU, Small kid)

2.3.2 Adjective

Adjectives agree with nouns and genders. A feminine adjective does not take plural marker while agreeing with a plural noun with feminine gender. For e.g.

Masculine singular

સારો છોકરો ('saro chhokaro' , Good Boy)

Masculine plural

સારા છોકરાઓ ('sara chhokarao' , Good Boys)

Feminine singular

સારી છોકરી ('sari chhokari' , Good girl)

Feminine plural

સારી છોકરીઓ ('sari chhokario' , Good girls)

2.3.3 Structure of verbs

Gujarati verbs have root+infinitive structure. Gujarati extends root verb to make causative sentence. For e.g.

જાદ પાડ્યુ. ('Zaad padyu' , A tree fell)

રામે જાદ પાડ્યુ. ('raame Zaad paadyu' , raam caused the tree fall)

કાને રામ પાસે જાદ પડાવ્યુ. ('kaane raam paase Zaad padaavyu' , Kan cause Ram who caused the tree fall)

3 Influence of other languages on Gujarati

As an Indo-Aryan language, Gujarati language is similar to Hindi, Marathi and Punjabi. Gujarati also has influence of other languages such as Urdu, English, Portuguese etc. Some examples are as follows,

• Urdu influence

દાવો (Urdu: dava English: Claim)

ફાયદો (Urdu: fayda English: Benefit)

• English influence

બેંક : Bank

ટેબલ : Table

• Portuguese influence

બટાટા: 'bataataa' potato

પાદરી: 'paadarI' father (Christian priest)

However, Similarity of Gujarati with Hindi helps in developing Gujarati WordNet using Hindi WordNet as the base. A brief comparison of Gujarati and Hindi is as follows,

• *Word Similarity* : Many words in Hindi and Gujarati originate from the same ancestor, so most of the concepts are described using same words in both the language.

• *Gender*: Gujarati language defines three genders while Hindi has only 2 genders.

• *Causative verbs*: Many languages differ in handling of causative verbs, which makes WordNets of languages difficult to link. However, both Hindi and Gujarati handle causative verbs in the same fashion.

Thus, Gujarati language has rich set of words derived from Indian languages as well as foreign languages. Gujarati language is similar to Hindi in structure and vocabulary.

4 Synset Development Approach

Gujarati WordNet is built using expansion approach (Vossen P., 1998). In this approach synsets are created by referring to the existing WordNet of the related language. Hindi is used as a source language to create synsets of Gujarati language. Due to this approach WordNet development process becomes faster as the gloss and synset of the source language is already available as a reference and it also becomes easier to link WordNets of two different languages.

The task of synset development for Gujarati language is further simplified by availability of the on line lexical resources like '*Bhagavad Go Mandal*' (Patel C. B.(ed) , 1958) and 'Gujarati Lexicon' (Chandaria R. , 2006). '*Bhagavad Go Mandal*' contains around 8.2 lacs words spread across 9 volumes. 'Gujarati Lexicon' is an another more recent effort. The on-line interface of Gujarati lexicon provides easy access to meanings, synonyms, antonyms, idioms, proverbs and phrases.

While we are using Hindi WordNet as a base, emphasis is given to understand the concept independently of a language and then to create a synset.

5 Synset Categorization

In some cases, there exists disagreement about a concept across languages. Many concepts of Hindi are not present in other languages or there is no indigenous lexeme for a concept in other language. So, to facilitate synset development using expansion approach, Hindi synsets are divided into following different categories,

- *Universal* : This set of concepts is present in all the languages and is essential and most frequently used. For e.g., 'सूर्य' (sun).
- *Pan-Indian* : This set of concepts is common in all Indian languages and linkable across all Indian languages but does not

have parallel concept in English. For example, 'तबला' (tabala)(An Indian rhythm instrument).

- *In-Family* : These are the concepts common in specific subsets of Indian languages and linkable across all languages of the family. For example: 'चाचा' (chacha)(paternal uncle) 'भतिजा' (bhatija) (brother's son)
- *Language Specific* : These concepts are specific to a language. It includes local food, festivals,etc. For example, 'ગરબા', ('garaba', *A form of dance performed to worship deity*) word is very specific to the state and the culture and does not appear in any other language.
- *Rare* : This includes very specific words adopted in most of the languages. It includes specific technical or scientific terms like, 'ngram'.
- *Synthesized* : These are the synsets created in a language due to the influence of other languages. These synsets are not natural to the language but needed to link synsets of two different languages.

Till date, 7163 universal synsets and 1356 Pan-Indian synsets have been manually identified and are now linked across all languages. Language specific and In-family synsets will be identified and linked by individual language group. Rare synsets will be adopted by languages. Synthesized synsets will be used to maintain a common index of all the synsets of Indian languages.

6 Synset Development status

Till date, 17653 synsets are built in the Gujarati WordNet. The category-wise count of synsets is as follows:

POS	No of Synsets
Noun	8683
Adjective	5753
Verb	2784
Adverb	433
Total	17653

This includes 7163 Universal and 1356 Pan-Indian synsets.

Gujarati WordNet is available online at '<http://www.cfilt.iitb.ac.in/gujarati>'.

7 Issues related to synset development

During the development of synsets, some disagreement was observed between Hindi concepts and Gujarati concepts.

7.1 Hindi synsets not linked with Gujarati

Many times it was observed that the concept present in Hindi was not present in Gujarati or even though the concept was present there was no indigenous lexeme for the concept. Following are some examples of Hindi synsets not linked with Gujarati,

- *Difference in concept description*

Concept: तुरही की तरह का एक बड़ा बाजा

(*TurahI kI taraha kaa Ek badaa baajaa*)

(*trumpet like A big musical instrument*)

(*A trumpet like big musical instrument*)

Synset: नरसिंहा, नरसिंगा, गोमुख

(*narasimhaa, narasimgaa, gomukha*)

No such concept is identified in Gujarati language. However, there is a concept in Gujarati language for a similar instrument which is used at war-front to announce beginning of a war.

- *No indigenous lexeme in Gujarati*

Concept: इत्र का व्यापर करनेवाला व्यक्ती

(*Itra kaa vyaapaara karanewala vyaktI*)

(*perfume's business person*)

(*A person who sells perfume.*)

Synset इत्र व्यापरी, इत्र फरोश, अत्तार

(*Itra vyaaparI, Itra Farosha, attaar*)

There is no indigenous lexeme for this concept in Gujarati language.

- *Difficult to adopt*

Concept: जो प्रवीष्ट न हुआ हो

(*jo pravisHT na huAa ho*)

(*who entered not is*)

(*one who is not entered.*)

Synset: अप्रवीष्ट

(*ApravishTa*)

Though this word can be translated in Gujarati language, it is not a native concept used in Gujarati language.

- *No such concept in Gujarati*

Concept: जो अकेला चरता या वीचरण करता हो
(*jo akelaa charataa ya vicharaN karata ho*)

(*who alone grazing or moving*)

(*one who is grazing or moving alone*)

Synset: पृथकचर

(*pruthakachara*)

(antonym of gregarious animal)

There is no such concept in Gujarati language.

7.2 Language specific synset

While major part of the day to day vocabulary of Gujarati is similar to that of Hindi, there are some concepts which are very specific to Gujarati language. These concepts are very specific to the culture of Gujarat. These concepts refer to food items, places, traditions, religion etc. Some examples are as follows:

- *Culture specific concept*

Concept : કોઈ ખાસ પ્રસંગે કસુંબો પીવા માટે બેગા થયું

(*koi khaasa prasange kasumbo pIvaa maaTe bhegaa thavu.*)

(*some special occasion 'kasumbo' drink get together*)

(*to get together to drink 'kasumbo' on some special occasion.*)

Synset : ડાયરો (ડાયરો, Daayaro)

- *Tradition specific concept*

Concept : એક ફળ કે જે લગ્ન પ્રસંગે વર કન્યા ના હાથે બાંધે છે

(*Ek faL ke je lagna prasange var kanya na haathe bandhe chhe*)

(*A fruit that marriage ceremony groom bride hand tie*)

(*A fruit that groom ties to the hand of bride in marriage ceremony.*)

Synset : મીઠળ (મીઠળ, mIMdhaL)

- *religion specific concept*

Concept : મોક્ષ માટે ભગવાન નું નામ લેતા લેતા ગીરનાર પર થી પડતું મુકવું.

(mokshHa maaTe bhagavaan nu naama leta leta giranaara para thI padatu mukavu.)

Synset : બૈરવજપ (ભૈરવજપ, bheiravajapa)

8 Validation plan

To verify and validate the Gujarat WordNet following guidelines are defined,

- *Gloss validation:* Synset should be defined in the context of its hypernymy. Gloss should also contain special attributes which distinguishes the concept from its co-hyponymy.
- *Synset completeness:* Synonymy set of the concept should cover all the words describing that concept. Other standardized lexical resource like, 'Bhagvad-Go-Mandal' and 'Gujarati lexicon' is referred to check synset completeness.
- *Lexical correctness:* 'saartha jodaNi kosh' is referred to check lexical correctness of the synset entry.
- *Part of speech validation:* Example sentences should be consistent with the part of speech of the synset.
- *WordNet coverage:* To check whether WordNet covers all the word of a language or not, available standardized lexical resources 'bhagvad-go-mandal' and 'Gujarati Lexicon' are referred.
- *User Feedback:* A feedback link is provided on Gujarati WordNet website. User feed back is used to improve quality of Gujarati WordNet.

9 Conclusion

Existence of Hindi WordNet and similarity between Hindi and Gujarati languages helped development of Gujarati WordNet. Synset categorization further simplified the synset linking process. It is observed that most of the top level concepts are common and easily linked. The concepts that vary across languages are

specific to culture and tradition. Most of these are noun concepts and do not have hyponymy. Many of these are singleton synsets that appear very low in the WordNet concept hierarchy. The future work is to identify and link language specific and in-family concepts. It is also required to develop lexical relations and to evaluate suitability of semantic relations of Hindi WordNet for Gujarati language.

Acknowledgements

Gujarati WordNet is developed as part of 'IndraDhanush - WordNets of seven Indian languages' project. The support of Ministry of Communication and Information Technology, Government of India is gratefully acknowledged.

References

- Christiane Fellbaum 1998. *WordNet: An Electronic Lexical Database*. MIT Press
- Piek Vossen 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers
- D. Chakrabarty P. Pande D. Narayan and P. Bhattacharyya 2002. *An experience in building the Indo WordNet - a WordNet for Hindi* International Conference on Global WordNet (GWC02), Mysore, India
- Patel C. B. 1958. *Bhagvad-Go-Mandal*. <http://www.bhagavadgomandalonline.com>.
- Ratilal Chandaria 2006. *Gujarati Lexicon* <http://www.gujaratilexicon.com>

Extending Czech WordNet Using a Bilingual Dictionary

Marek Blahuš

Masaryk University

Brno, Czech Republic

gwc2012@blahus.cz

Karel Pala

Masaryk University

Brno, Czech Republic

pala@fi.muni.cz

Abstract

In this paper we describe semi-automatical extending of the Czech WordNet lexical database (48,000 literals in 28,000 synsets) by translation of English literals from existing synsets in Princeton WordNet. We make use of a machine-readable bilingual dictionary to extract English-Czech translation pairs, search the English literals in Princeton WordNet and in case of a high-confidence match we transfer the literal into Czech WordNet. Along with literals, new synsets parallel to the English ones and identified by ILI are introduced into CzechWordNet, including information on their ILR (Internal Language Relations) such as hypernymy/hyponymy. The paper describes the parsing of the dictionary data, extraction of translation pairs and the criteria used for estimating the confidence level of a match. Results of the work are 36,228 added literals and 12,403 created synsets. An overview of previous similar attempts for other languages is also included.

1 Credits

This paper is an adaption of a Master's thesis defended by Marek Blahuš at Masaryk University in June 2011, under the supervision of Karel Pala (Blahuš, 2011).

2 Introduction

WordNet (developed at Princeton University by Miller (1995) and Fellbaum (1998), further referred to as PWN), a lexical database for the English language, describes a large amount of semantic relations among concepts. If an approach is found to automatize the translation of synsets

(synonym sets) by using a suitable linguistic resource as a reference, e.g. a bilingual dictionary, PWN's data might be used to stimulate the growth of its other-language counterparts.

In the summer of 2010, we acquired machine-readable data from the largest one-volume English-Czech dictionary ever published, which has prepared the ground for this attempt at automatically suggesting extensions to the Czech WordNet.

2.1 Czech WordNet

The unique concept introduced by PWN and the interest in the range of its applications in Natural Language Processing had stirred up an international research cooperation which eventually culminated in the founding of the Global WordNet Association (2000). The main initiative came from the members of the EuroWordNet project (Vossen, 1997) that ran from 1996 to 1999. It has produced lexical databases of WordNet type for seven European languages, including Czech. BalkaNet (Tufi et al., 2004), a later project running from 2001 to 2004, has contributed by WordNets for five Balkan languages. A common link among all these databases is the Inter-Lingual Index (hereafter ILI) – a unique assignment of identification numbers to all synsets of PWN which makes it possible to establish links among PWN synsets and their equivalents in any other language.

The CzechWordNet (Pala and Ševeček, 1999) (hereafter CzWN) has been developed at the Faculty of Informatics of Masaryk University since 1998. During the second phase of EuroWordNet, approximately 14,000 synsets were gathered that formed the first version of CzWN in 1999. Before extending CzWN as described below, in 2011, it comprised 34,026 literals organized in 28,478 synsets for a total of 47,542 word-sense pairs. Out of this, 21,018 (74 %) were noun synsets, 5162 (18 %) were verb synsets, 2129 (7 %) were adjec-

tive synsets and 166 (1 %) were adverb synsets. The ILI indices that link CzWN synsets to PWN are those of PWN 2.0 (remapping them to PWN 3.0 has been considered but not done yet). VerbaLex (Hlaváčková, 2008) is a database of verb valency frames created since 2005 at the Faculty of Informatics, which has also been linked by ILI to PWN 2.0, is large in size (approximately 20,000 valency frames) and may eventually replace the verbal part of CzWN.

2.2 Comprehensive English-Czech Dictionary

The Comprehensive English-Czech Dictionary (hereafter referred to by its Czech abbreviation VAC) compiled by Josef Fronek (2006) is the largest one-volume English-Czech dictionary ever published. It contains more than 100,000 headwords and sub-headwords, more than 200,000 words and phrases and roughly 400,000 equivalents.

3 Overview of Previous WordNet Translation Attempts

In the foreword of (Fellbaum, 1998), the creators of PWN mention that they have used various dictionaries, word lists, corpora and thesauri as sources for building their vocabulary over the time. There have been many attempts at using bilingual (and multilingual) dictionaries for the construction or enlargement of non-English WordNets. In this paper we provide a short summary for two of them that have inspired our attempt the most.

MLSN (Cook, 2008), the multi-lingual semantic network, started in 2005, has created a database consisting of words and relations among them, plus a web interface to enable the public to contribute data. In MLSN's algorithm, freely available bilingual dictionaries are used to automatically translate entries. All monosemous English nouns from PWN for which there is a single translation in the target language are translated back using a complementary dictionary (target language to English). The result of this counter-checking is a set of one or more English literals. If it equals exactly to the original PWN synset of the English noun ("high confidence"), or if at least some of this synset's members (and at the same time no non-member literals) have been discovered ("medium confidence"), the result gets im-

ported into the MLSN database.

The authors report a high reliability, although at the cost of excluding many common words that are polysemous. Between eight and seventeen thousand literals could be imported by this algorithm for Japanese, German and Chinese. Moreover, inclusion of a second independent dictionary – Wiktionary (Wiktionary authors, 2011) – has further improved the quality of the results. Manual evaluation has shown that 95–97 % of the high-confidence results and 94–98 % of the medium-confidence results were achieved.

WordNet Builder (Sathapornrungkij and Pluempiwiriyawej, 2005) is a Thai project that has designed a WordNet-building system with a clearly defined architecture. Its two main components are a Machine-Readable Dictionary Extractor and a WordNet Constructor that communicate with each other through a Link Analyzer. Three kinds of links are distinguished: *translation links* (between English and Thai words), *semantic links* (between English words and their meaning) and *candidate links* (between Thai words and their meaning). A set of 13 criteria, divided into three groups (monosemic, polysemic and structural), has been used. For each, a sample of translation links was manually verified and a statistical classification model was constructed which was later used to classify and validate the remaining ones. WordNet Builder is very instructive by the clear distinction of the possible word-to-word-to-synset-mapping situations (as expressed by the criteria) which are depicted in the form of diagrams.

With little surprise, the best criterion (92 % correctness) was the "monosemic one-to-one" criterion. It considers only those monosemous PWN literals that have a single translation equivalent in Thai. The worst criterion (49.25 % correctness) was "polysemic many-to-many".

4 Proposed Algorithm

Due to the simplicity and effectiveness of MLSN's algorithm, we drew inspiration particularly from it, but we made modifications to it and expressed it in terms of the criteria introduced by WordNet Builder. In contrast with MLSN, we consider not only the monosemous English literals that have a single Czech translation in VAC (criterion 1: "monosemic one-to-one"), but also those that have several translation equivalents in the dictio-

nary (criterion 3: "monosemic many-to-one"). On the other hand, since we do not have a machine-readable Czech-English dictionary of similar size at our disposal, we do not translate the found equivalents back to English to compare the results with the original synset as MLSN does.

Instead, we benefit from a special property of VAC, namely its design as an active dictionary, in contrast to a passive (i.e. typically translational) dictionary. This has the consequence that individual senses of a single English word are being carefully distinguished and identified by Arabic numerals. Sometimes, this distinction goes even into three levels (subsenses), using also Roman numerals and upper-case letters.

The research we have done has confirmed the validity of the above statement. Average polysemy of English literals in VAC has been found same or higher in comparison with polysemy in PWN 3.0 (Princeton University, 2011). On average, there are:

- 1.51 senses per noun in VAC as compared to only 1.24 in PWN 3.0;
- 2.22 senses per verb as compared to 2.17;
- 1.36 per adjective as compared to 1.40;
- 1.33 per adjective as compared to 1.25.

This observation makes us believe that the sense distinctions made in VAC are at least as fine-grained as in PWN and that therefore, if we restrict our attention only to literals that are monosemous in both PWN and VAC, we can be confident enough to assume that we indeed work with the same sense on both sides.

We propose the following algorithm (in pseudocode), divided into two main blocks that we describe in the sections below:

1. Dictionary Data Parser

- (a) Transform VAC data from presentational markup to descriptive markup.
- (b) Extract translation pairs from VAC data.

2. WordNet Synset Generator

- (a) Keep only pairs in which English literals are monosemous.
- (b) If desired, keep only pairs with unique source literals (one-to-one translations).
- (c) Match VAC English literals with monosemous PWN literals.

- (d) See which matched synsets are present in CzWN and which not.
- (e) Recalculate and transfer ILR to new CzWN synsets.
- (f) Merge new synsets and new literals with CzWN.

5 Dictionary Data Parser

Because VAC, source of our dictionary data, had never been used in NLP before, we have spent a considerable amount of time by parsing the acquired data so that they would suit to our needs and be in a form that can be passed to a WordNet constructor.

5.1 Transformation of Dictionary Data

The dictionary came in form of a huge XML file with most of the markup obviously designed to serve presentational purposes rather than showing the structure of entries. Fortunately, we were able to reconstruct the semantic structure of dictionary entries by taking notice of distinctive font and punctuation patterns. Still, some inaccuracy could not be avoided, because of ambiguities in markup design and its occasional inconsistent use. Treatment of syntactic exceptions has eventually consisted in a significant part of the work spent on transforming the dictionary file into a more descriptive markup.

In VAC, each dictionary entry is encapsulated in its own XML element. These elements are read one-by-one from the XML file – represented using the DOM formalism (Hors et al., 2007) – and each is processed on its own by means of an XSL transformation run in the Saxon XSLT processor (Kay, 2010).

By searching for specific elements, each entry is structured into head sections (containing headwords) and body sections (each contains information specific to a single sense of the most recent headword). Again, because of occasional markup inconsistencies, rather complicated matching expressions have had to be used; all treated exceptions have been illustrated by examples in the source code.

In the first head section of a dictionary entry, we look for headwords, their spelling variants or alternatives, as well as grammatical marks (part of speech) and references to other headwords ("see also"). In the first body section, we go down the sense hierarchy, noting all sense numbers found

(to recognize polysemy), grammatical marks (part of speech of a sense may differ from the headword), all translation equivalents (of the most recent headword), all examples and idioms and their translations. All other kinds of information (such as pronunciation) get ignored. In eventual subsequent head sections, new headwords (such as phrasal verbs) are defined as derivations from the first headword of the dictionary entry, for instance using the symbol.

5.2 Extracting Translation Pairs

To extract translation pairs, we need contextual (as opposed to purely hierarchical) information on processed XML elements, but XSLT transformations are stateless. We solve this by adding another XSLT stylesheet, processed during the event-driven XML parser SAX (Megginson, 2004), which scans elements in document order and for each translation pair, a five-tuple is output that consists of: the respective pair of literals, the sense number of its English constituent (so that we may still distinguish pairs originating from polysemous headwords), information on part of speech (defined as the value of the most recently encountered part of speech mark within the same entry), and the type of that pair (either headword translation equivalent, example or idiom).

English literals (headwords and their derivatives, phrasal verbs, examples, and idioms) are held in a hierarchy that changes progressively while we scan through an entry so that it always describes the current context. Every time we reach a Czech literal, we match it with the most recently encountered English literal and output the two as a translation pair. If more English literals are grouped (e.g. several spellings variations of the same headword), we output a five-tuple for each combination.

Special attention has been paid to the treatment of alternatives (expressed by a slash or OR) and parentheses that are vastly applied in both English and Czech literals to make dictionary entries shorter. We try to expand each such expression to all the different values it represents. Note that possible overproduction of English literals is not a major problem, because in most cases those are non-existent in PWN and therefore get ignored.

Our expansion algorithm produces satisfactory results for most cases, but it fails in the difficult ones, from which some are ambiguous even to

a human. Additionally, the algorithm performs also expansion of headwords, which are usually referred to only by or by their first letter and a period. The basic rules of the algorithm are:

- If the number of slash-delimited parts is the same for both the source and the target literals, split the literals at the slashes and generate new translation pairs from each pair of such parts (due to the similarity to the mathematical operation, call this a *dot product*).
- If the number of slash-delimited parts differs, create new translation pairs for each possible combination of a source literal's part and a target's literal part (call this a *cross product*).
- For any translation pair containing an OR, perform *cross product* with taking OR as delimiter.
- For any translation pair containing a parenthetical expression, generate all possible expansions for each of the two literals by either including or excluding each of the parentheses' content (call the result a *power set*) and eventually perform a *cross product* on the resulting two *power sets*.
- When performing a *power set*, run this algorithm recursively on the content of parentheses in case it contains any slashes or ORs.

6 WordNet Synset Generator

We use an XML file structure defined by VisDic (Horák and Smrž, 2003) to represent all our WordNet data (PWN, CzWN), due to the possibility to parse such files with existing generic parsers. In it, each SYNSET element contains an ID element (its identifier in ILI), a POS element (part of speech) and one or more LITERALS that form the SYNSET. If there are several LITERAL elements of identical value in a WordNet file, each such occurrence must be distinguished by a unique SENSE number.

We work only with translation pairs that could normally appear in a WordNet, i.e. we exclude idioms and examples and we limit us to nouns, verbs, adjectives and adverbs. Only those English literals that are monosemous both in PWN and VAC are kept. A literal in PWN is monosemous if it appears only once in the whole database. A literal is monosemous in VAC if it appears with a unique sense number across all output translation

pairs. Dummy sense numbers are introduced if a headword did not produce any translation pair. Better results (at the expense of decreasing quantity) may be achieved by considering only English literals having a single Czech translation equivalent (step 2b of the pseudocode above and a move from WordNet Builder's criterion 3 to criterion 1).

With help of ILI, we search CzWN for Czech equivalent of the PWN synset corresponding to each VAC literal. If the results is negative, we establish a new synset in CzWN, otherwise we only add any VAC literals new to this synset and call this a changed synset. Newly added literals may be marked by prepending "X" to make their identification easier during later inspection in VisDic.

7 Evaluation

The extended version of CzWN, created as described above, now contains 83,769 literals (growth by 76 %) organized into 40,621 synsets (growth by 43 %). The number of unique literals is 63,775, which is an average polysemy of 1.31 senses per literal (monosemy has increased). There are on average 2.06 literals in a synset (synset size has grown due to augmentation). Out of the synsets, 27,658 are noun synsets (increase by 6640, or 31.6 %), 5852 are verb synsets (increase by 690, or 13.3 %), 5651 are adjective synsets (increase by 3522, or 165.4 %) and 1457 are adverb synsets (increase by 1291, or 877.7 %).

VAC data contains 54,046 entries, from which we were able to generate 315,991 translation pairs. Out of these, 194,276 are headword or phrasal verb equivalents, 111,774 are examples and 9941 are idioms. Out of the headword or phrasal verb equivalents, 47,588 (24.5 %) are monosemous. Out of these, 21,700 have a single Czech translation ("one-to-one monosemy").

Because of unsupervised nature of the extension, the newly produced CzWN data need to be inspected manually. An experimental manual assessment of a hundred of synsets generated with the "one-to-one" limitation switched on has identified that two synsets in a hundred had to be reconsidered by a human editor and one needed a minor change (result of incorrect slash expansion). This indicates a satisfactory 97–98 % reliability rate of the results. There are in total 7399 synsets generated with such a high confidence. Further analysis of the reliability of the results as well as inspection of the newly added and modified literals have not

yet been performed but will be done by linguists who had previously helped producing CzWN data manually.

As a side effect, parsing of the VAC dictionary has been of direct use to PRESEMT (2011), a running machine translation project.

8 Conclusion

We have designed and implemented a set of tools that is capable of automatically generating and augmenting CzWN synsets, based on analysis of English synsets in PWN and of English-Czech equivalents extracted from VAC data transformed from presentational to descriptive markup.

We have shown that by searching PWN for English literals from VAC, links may be established that make it possible to create new synsets in CzWN, using the Czech translations of these literals to populate the synsets, and the synsets may be linked to their PWN counterparts by means of ILI (Inter-Lingual Index). Along with this linking, also ILR (Internal Language Relations) may be transferred from English synsets into the new Czech ones, on demand, including the possibility of automatically creating a hypernymy/hyponymy relation between the new Czech synset and its closest existing Czech hypernym, by tracing the hypernym axis in PWN and looking for synsets that already have a Czech equivalent.

In spite of the fact that only a subset of the whole dictionary could be used, due to the constraint that only monosemous literals were to be taken into account, we have managed to almost double the size of CzWN in the number of literals and increase it by a half in the number of synsets. First observations suggest that the quality of the produced data is high enough for it to be passed to linguists who would check all the new and modified synsets and make any necessary corrections.

9 Future Directions

Future directions include trying to match VAC senses with PWN senses to overcome polysemy; identifying the most sustainable way of merging CzWN data and VerbaLex data; remapping CzWN from PWN 2.0 ILI to PWN 3.0 ILI.

Acknowledgments

The work described in this paper was funded by the Czech Ministry of Education under the projects LC536 and Lindat-Clarin.

References

- BLAHUŠ, Marek. *Extending Czech WordNet Using a Bilingual Dictionary*. Master's thesis, Brno : Masarykova univerzita, Fakulta informatiky, 2011, 42 pp., <http://is.muni.cz/th/172464/fi_m/>
- COOK, Darren. *MLSN: A multi-lingual semantic network*. In: 14th Annual Meeting of the Association for Natural Language Processing. Tokyo, 2008.
- FELLBAUM, Christiane. *WordNet : an electronic lexical database*. Cambridge, Mass : MIT Press, c1998, xxii, 423 pp., 9780262061971.
- FRONEK, Josef. *Velký anglicko-český slovník [Comprehensive English-Czech Dictionary]*. Vydání 1., Praha : LEDA, 2006, 1734 pp., 80-7335-022-X.
- HLAVÁČKOVÁ, Dana. *Databáze slovesných valenčních rámci VerbaLex*. Ph.D. thesis, Brno : Masarykova univerzita, Filozofická fakulta, Ústav českého jazyka, 2008, 136 pp., <http://is.muni.cz/th/17907/ff_d/>
- HORÁK, Aleš and SMRŽ, Pavel. *Visdic – WordNet browsing and editing tool*. In: Proceedings of the Second International WordNet Conference – GWC 2004. Brno : Masaryk University, pp. 136–141, 2003.
- HORS, Arnaud Le and HÉGARET, Philippe Le and WOOD, Lauren and NICOL, Gavin and ROBIE, Jonathan and CHAMPION, Mike and BYRNE, Steve. *Document Object Model (DOM) Level 2 Core Specification*. Version 1.0, W3C Recommendation 13 November, 2000, <<http://www.w3.org/TR/DOM-Level-2-Core/>>
- PRESEMT (Pattern REcognition-based Statistically Enhanced MT). Athens : Institute for Language & Speech Processing [coordinator], running project, <<http://www.presemt.eu/>>
- KAY, Michael H. *Saxon: The XSLT and XQuery Processor*. Version 9.3, 30 October 2010, <<http://saxon.sourceforge.net/>>
- MEGGINSON, David. *SAX: Simple API for XML*. Version 2.0.2, 27 April 2004, <<http://www.saxproject.org/>>
- MILLER, George A. *WordNet: A Lexical Database for English*. In: Communications of the ACM. New York : ACM Press, 1995, Vol. 38, No. 11, pp. 39–41, ISSN 0001-0782.
- PALA, Karel and ŠEVEČEK, Pavel. *The Czech WordNet, final report*. Brno : Masarykova univerzita, 1999, 21 pp., technical report, <<http://www.illc.uva.nl/EuroWordNet/docs/CzechWordnetPS.zip>>
- wnstats - WordNet 3.0 database statistics [online]. In: WordNet 3.0 Reference Manual. Princeton University, accessed 2011-05-28, <<http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html>>
- SATHAPORNRUNGKIJ, Patanakul and PLUEMPI-TIWIRIYAWEJ, Charnyote. *Construction of Thai WordNet Lexical Database from Machine Readable Dictionaries*. In: Conference Proceedings: the tenth Machine Translation Summit. Thailand, 2005, pp. 87–92.
- TUFI, Dan and CRISTEA, Dan and STAMOU, Sofia. *BalkaNet: Aims, methods, results and perspectives; a general overview*. In: Romanian Journal of Information Science and Technology. Bucureti : Romanian Academy, 2004, Vol. 7, No. 1–2, pp. 9–43, ISSN 1453-8245.
- VOSSEN, Piek. *EuroWordNet: a multilingual database for information retrieval*. In: Proceedings of DELOS workshop on Cross-language Information Retrieval. Zrich, 1997, pp. 85–94.
- Wiktionary: *The Free Dictionary*. Dictionary online, Wikimedia Foundation Inc., <<http://en.wiktionary.org/>> .

Japanese SemCor: A Sense-tagged Corpus of Japanese

Francis Bond,^{*,***} Timothy Baldwin,^{**} Richard Fothergill^{**} and Kiyotaka Uchimoto^{***}

^{*} Linguistics and Multilingual Studies, Nanyang Technological University, Singapore

^{**} Computer Science and Software Engineering, Melbourne University, Australia

^{***} National Institute of Information and Communications Technology, Japan

bond@ieee.org, tb@ldwin.net, richard.fothergill@gmail.com, uchimoto@nict.go.jp

Abstract

In this paper we describe the creation of the Japanese SemCor (JSEMCor) sense-tagged corpus of Japanese. The corpus is a translation of the English SEMCOR, with senses projected across from English. The final corpus consists of 14,169 sentences with 150,555 content words of which 58,265 are sense tagged. The corpus is one of the corpora used to provide sense frequency data for the Japanese Wordnet.

1 Introduction

Wordnets have been shown to have utility across a broad range of applications, largely in combination with sense frequency data and sense-tagged corpora. This paper describes Japanese SemCor (JSEMCor), a sense-tagged corpus for the Japanese Wordnet (Isahara et al., 2008), based on translation of the English SEMCOR and sense projection.

In order to produce annotated text quickly and cheaply, we adopt the method of **annotation transfer** pioneered in the Italian MULTISEMCOR (Bentivogli and Pianta, 2005). In this approach, a sense-tagged text in one language is translated into the language in question, and the sense annotations from the original corpus are projected onto the new language. The sense projection is based on a wordnet in the target language which is aligned with the wordnet that was used to sense tag the source language text. Bentivogli and Pianta (2005) found that annotation transfer led to sense tagging with a precision of 86% and coverage of 81% (that is 19% of open class words still needed to be annotated), at less cost than annotating from scratch. The main differences in our case are: (1) the target language (Japanese) is linguistically further removed from the source language (English);

and (2) to boost coverage, we provide the translators with sense-specific translations of each open class words to optionally include in their translations.

Similarly to MULTISEMCOR, our method takes English SEMCOR and translates it into the target language. In addition to the immediate objective of deriving a sense-tagged corpus of Japanese based on Japanese Wordnet, we also create a trilingual (English–Italian–Japanese) sensebank, with potential applications in other tasks such as translation. Because the Japanese (and Italian) texts are translated, the sense distribution may not be truly representative of native Japanese text. Ultimately, we aim to supplement JSEMCor with other sense-tagged data, based on native Japanese text.

Finally, in the same way that the English and Italian annotation revealed missing word senses for their respective wordnets, we expect to find and correct such errors in the Japanese Wordnet, which will be fed back to the developers.

In the next section we give a brief description of the base resources used in the creation of JSEMCor. Next, we describe the creation, size and distribution of JSEMCor (§3). Finally, we discuss future work (§4) and then conclude.

2 Resources

2.1 The English Wordnet

The wordnet used both to tag the English SEMCOR corpus (§2.3) and as the backbone of the Japanese wordnet (§2.2) is the Princeton WordNet of English (Fellbaum, 1998). SEMCOR is tagged with tags from version 1.6 and the Japanese wordnet aligns with version 3.0. PWN has a rich structure of semantic relations, but we are only using it as a source of sense inventories in this task.

2.2 The Japanese Wordnet

The Japanese Wordnet is a large scale, freely available, semantic dictionary of Japanese. The National Institute of Information and Communications Technology (NICT) started developing the Japanese Wordnet in 2006, as part of its support for Natural Language Processing research in Japan. The first version (0.9) was released in February 2009. In the initial phase Japanese equivalents were added to synsets of the Princeton WordNet. These have been expanded and corrected in subsequent releases. The current release is version 1.1. It contains 57,238 synsets (concepts), 93,834 unique Japanese words and 158,058 senses (synset–word pairs). All synsets have Japanese definitions, and over 45,000 also have examples.

We give an example of an entry in Figure 1.

From the beginning, the Japanese Wordnet project planned to tag text in order to verify its coverage and get distribution information (Isahara et al., 2008), but no tagged text has been released so far.

2.3 SemCor and MultiSemCor

The English SEMCOR corpus is a sense-tagged corpus of English created at Princeton University by the WordNet Project research team (Landes et al., 1998). It was created very early in the WordNet project, and was one of the first sense-tagged corpora produced for any language. The corpus consists of a subset of the Brown Corpus (Francis and Kucera, 1979), and has been part-of-speech tagged and sense tagged. We use the subset of SEMCOR which was translated into Italian as part of MULTISEMCOR 1.1 (Bentivogli and Pianta, 2005).

MULTISEMCOR is an English/Italian parallel corpus created by translating the English SEMCOR corpus into Italian. Texts are aligned at the sentence and word level, and annotated with part of speech, lemma and word sense (PWN 1.6). MULTISEMCOR version 1.1 contains 116 English texts: 14,144 sentences and 261,283 tokens, of which 119,802 tokens are annotated with senses. These are aligned with their corresponding Italian translations. In this paper we only use the English texts, which are freely available.¹ The MULTISEMCOR team reports tag errors for around 2.5% of the English open-class tokens in the English SEMCOR (Bentivogli and Pianta, 2005).

3 Japanese SemCor

The initial data for the translation was created by taking the English SEMCOR data and mapping the senses to version 3.0 using the mappings created by Daude et al. (2003). These senses were then used to look up synsets in the Japanese Wordnet to be presented to the translators.

3.1 Creation

Similar to MULTISEMCOR, sense annotation in JSEMCor was set up as a translation task, where translators were provided with a SEMCOR sentence in English and asked to generate a Japanese translation using the interface depicted in Figure 2. For each sense-indexed word in the original SEMCOR data, we provided translators with a list of all words contained in the corresponding Japanese Wordnet synset. Clicking on one of these words both appended that lemma to the translation, and recorded clickthrough data for the word, which

¹The Italian texts are available free for research from the Istituto Trentino Di Cultura (ITC) <http://multisemcor.itc.it>.

Synset	02076196-n
Synonyms	[ja 海豹, アザラシ, シール en seal]
Def (en)	“any of numerous marine mammals that come on shore to breed; chiefly of cold regions”
Def (ja)	「繁殖のために岸に上がる海洋性哺乳動物の各種；主に寒帯地域に」
Hypernyms	アシカ亜目/pinniped
Hyponyms	?/crabeater_seal ?/eared_seal 海鼴/earless_seal


animal/seal.png

Figure 1: Example Entry for Seal/海豹

Last 10 translations:

128) 悪いことをした珍・ | Go Translator history | Document view | All data

Hold Unsure

Prev Remove temptations .
誘惑を取りはらうこと。
English: Remove the child from the scene of his misbehavior .
Japanese: 悪いことをした現場から子供を離す。
Clear **Save**

remove: 除去 除する 除す 除ける 除く 跳ね除ける 撥ね除ける 撤収 撤去 排除 持ち去る 抜去る 抜き去る 抜きさる 扣除 扣除ける 扣い除ける 引退ける 引去る 引上げる 引き退ける 引き去る 奪取る 奪い取る 奪い去る 取除く 取捨てる 取扱う 取外す 取去る 取上げる 取る 取り除ける 取り除く 取り捨てる 取り扱う 取り去る 取り上げる 取りはらう 取りはずす 取りのぞく 取りのける 取りする 取りさる 取りあげる 取っ払う 取っ外す 取っぱらう 取っぱずす 収去 すぐ

child: 子供 愛児 子種 子弟 子女 子ども 子 お子様 お子さん

scene: 現場 現地 所 場所 場

misbehavior: 非行 非 酔行 罪 汚行 暴 悪行 心得違い咎 不行跡 不始末 不埒 不埒 不仕末

Comments:

Figure 2: Screen shot of the annotation interface, for the SEMCOR sentence *Remove the child from the scene of his misbehaviour*

provided the basis for the ultimate sense tags in the translation. Translators also had the option to leave a comment (e.g. if they wanted to note something about the translation lists, or the source English), to mark their lack of confidence in a translation (via the “Unsure” checkbox), or to leave a translation to come back to (via the “Hold” checkbox).

In the example in Figure 2, shown in (1), the translator has used the words 子供 *kodomo* “child” and 現場 *geNba* “scene”, but has not made use of any of the translations for *remove* or *misbehavior* in their final translation (as indicated both by the lack of a highlighted translation and the ticked checkboxes for the respective words). They had the option of adding the new translation of *remove* to the synset, but did not use it here. The single word *misbehavior* was translated as the multi-word expression 悪いことをした *warui-koto-*

wo shita “did something bad”, probably because all of the translations of *misbehavior* sound more criminal than their English equivalents, making them inappropriate for child behaviour.

- (1) a. *Remove^a the child^b₁ from the scene^c₁ of his misbehavior₁.*
- b. 悪い? こと? を した? 子供^b₁
 warui koto wo shita kodomo
 bad thing ACC done child
 を 現場^c₁ から 離す^a₁ 。
 wo genba kara hanasu .
 ACC scene from remove .

The final result for this sentence is that, of the four sense-tagged words in the original, two words have their sense transferred (^{b,c}), one word could be transferred if the new lemma were added to it (^a), and one word gets translated into three, none of which can be easily linked.

At the outset of the translation process, sentences were allocated to translators from a global *sentence* queue, meaning that if two translators were working in tandem, a given translator would often not translate contiguous sentences from a given document, potentially leading to lack of coherence in the translations. While translation coherence was not of primary interest, we switched across to allocating data from a *document* queue about 20% of the way through the translation process, in response to concerns over the resultant consistency in sense annotations within a given document, and requests from the translators. As part of this, we provided support for a “Document view”, to allow the translator to look over a document in its entirety, including whatever progress had been made through the translation. We also gave translators the option of viewing the source English and translation for the immediately preceding sentence (*Remove temptations* and 疑惑を払うこと *giwaku-o harau koto*, resp., in Figure 2). They could also view their past 10 translations via the pulldown menu at the top-left of the translation page.

Translators were instructed to use translations provided in the list where possible, in order to maximise sense tagging coverage, except where this led to stilted Japanese (e.g. in translating an English deictic pronoun literally, rather than using a zero pronoun). The purpose of translation lists and the need for the clickthrough data was explained to the translators, although none of the translators were computational linguists, so the significance of sense tagging and the resulting sense-tagged corpus wasn’t self-evident to them. Translators were also instructed to:

- use formal “editorial” Japanese, e.g. using the である *dearu* form of the copular, unless the text was clearly written in a colloquial or other style;
- attempt to determine the canonical translation/transliteration of proper names where possible, and failing this, to transliterate, flagging the translation as “Hold” if unsure of the pronunciation; acronyms were to be left as is, unless there was a well-known Japanese rendering of the acronym (e.g. *METI* for 通産省 *tsūsaNshō*);
- refrain from including alternative translations, e.g. in parentheses, in cases of doubt;

Corpus	SEMCOR	JSEMCOR
Sentences	12,842	14,169
Words	261,283	382,762
Content Words	119,802	150,555

Table 1: Corpus Size

- be faithful to the English sentence tokenisation (i.e. never translate multiple English sentences into a single Japanese translation), but to translate into multiple sentences in cases where it improved readability (e.g. for particularly long or heavily embedded English sentences);
- reorder the words where necessary to maximise readability in Japanese (esp. for conjunctions of nouns or adjectives);
- include discourse connectives where it improved overall sentence and document readability, irrespective of whether a corresponding sentential adverb (or equivalent) was included in the original English sentence.

3.2 Statistics

In contrast to Bentivogli and Pianta (2005), we have used manual rather than automatic word alignment. However, the alignment requires some post-processing before annotation transfer can occur. In this section, we look at various statistics of the alignment and annotation transfer process.

The word-alignment clickthrough data produced by our translators maps tokens in SEMCOR to lemmas in Japanese Wordnet, within the context of a translated sentence. In the following, we refer to a translated lemma in context as a **translation lemma**. Each translation lemma must be mapped onto the text of the translated sentence to complete the word alignment.

We perform this mapping automatically by first tokenising the sentence with the morphological analyser MeCab using the IPAAdic lexicon and tagset (Kudo et al., 2004) and using the part of speech and lemma information it provides. This results in 382,762 tokens overall and 148,249 open class tokens, giving averages of 27 and 10.5 per sentence respectively.

The segmentation MeCab produces is fine grained relative to both English and to the Japanese Wordnet — in particular splitting compounds into their components — so we map trans-

lation lemmas to sequences of tokens. We accept a sequence of tokens as a match for a Princeton WordNet lemma if all parts in the translation match in their canonical word order, optionally allowing the final token to be in its lemmatised form, which is a convenient heuristic for lemmatising Japanese compounds. The numbers are summarized in Table 1.

Of 61,827 translation lemmas available, 7,551 are compounds with respect to IPAdic. Of the rest, 44,813 are single token and 9,463 are not found in the translation: the translation interface allowed free editing of the translation text but did not allow clickthrough word alignments to be undone.

Note that the resulting word alignment is not one-to-one: 1,734 translation lemmas come from more than one source word, though only 190 come from more than one source lemma (and none from more than two). Conversely, 3,252 translation lemmas match more than once in the translated sentence. Also, the alignment coverage is not complete: 51,450 sense tagged tokens in SEMCOR have not been translated, and 90,525 open class tokens in the Japanese sentence translations have no translation lemma mapped to them. Part of speech distributions for unaligned tokens in both languages are shown in Table 2.

After completion of the word alignment, we perform the annotation transfer. For a number of reasons, annotation transfer can result in zero or multiple senses being assigned to a word-aligned translation:

- Due to rearrangement of senses between WordNet versions 1.6 and 3.0, some SEMCOR tokens are annotated with deleted senses and others with more than one sense.
- We introduce additional variation in annotation multiplicity with a many-to-many word alignment.
- The presence in the translation data of user-contributed word translations means that an aligned word is not always in the transferred synset in Japanese Wordnet. In fact, this occurs 13,857 times, suggesting a large number of potential new synset memberships for Japanese Wordnet.

Therefore, of the 61,827 translation lemmas, 131 are assigned more than one sense and 13,771 have none. The remaining 47,925 translation lemmas

are assigned a single sense. After taking into account translation lemmas which appear more than once — or not at all — in the target sentence, 46,121 words receive tags from the annotation transfer.

Due to the granularity mismatch between IPAdic and Japanese Wordnet, we take the additional step of mapping Japanese Wordnet lemmas to portions of text without word-aligned translations. The resulting compounds (or single tokens) do not receive a sense tag but are annotated with Japanese Wordnet lemma and part of speech. Where potential matches overlap, precedence is given first to longer matches (e.g., 米国政府 *beiko-kuseifu* “Washington” is chosen over 政府 *seifu* “government”) and then to earlier matches (e.g. 近代化 *kiNdaika* “modernisation” is chosen over 化する *ka-suru* “to change” where they intersect in the out-of-vocabulary 近代化する *kiNdaika-suru* “to modernise”). This process produces an additional 61,495 unaligned words. We then include open class MeCab tokens which have still not been assigned a Japanese Wordnet lemma as an additional 34,329 words.

Finally, there are 12,144 monosemous Japanese words (with only a single sense) which were not annotated in the translation process, either because sense transfer fails or because the word is not aligned. Applying these single sense annotations brings the total number of sense annotated words to 58,265.

3.3 Distribution

We use the Kyoto Annotation Format (KAF) to share the corpus (Bosma et al., 2009). This is an emerging standard for wordnet annotation. We only use the two lowest layers (text and term), not including any higher levels such as dependencies or geodata. In order to make the data accessible, we will release it under the same license as the English SEMCOR. JSEMCOR is distributed with the Japanese Wordnet, available from <http://nlpwww.nict.go.jp/wn-ja/>.

A sample KAF record is presented in Figure 3, containing two words with Japanese Wordnet senses (学校 *gakkō* “school” and 戻る *modoru* “return”), IPAdic part-of-speech tags for all tokens, and file and sentence IDs which align with English SEMCOR.

Part of Speech	English Tokens	Japanese Tokens
Verb	13,457	24,698
Noun	9,979	41,394
Adjective	10,337	2,794
Adverb	12,321	5,635

Table 2: Part of speech distribution for tokens without word alignment

4 Discussion and Future Work

We were able to transfer far fewer senses than the MULTISEMCOR (39% vs. 81%). One major reason for this is that the missing terms that this annotation project has found have not yet been added to the Japanese Wordnet. Adding them will raise the coverage by another 9%. Another reason is that we are currently overcounting untagged senses — if a word should be tagged as a multiword expression we count it once as the MWE and once for each of the single terms. However, the greatest reason is the fundamental differences between Japanese and English. There were three major causes that made transfer impossible. The first is that in many cases a word-for-word translation is unnatural — either there is a lexical gap in Japanese so that the English term does not have any translation, or the direct translation has a different connotation.

A major cause of lexical gaps is part-of-speech mismatches. For example, the English Wordnet has these three entries for *French*:² French_n_1 “a native or inhabitant of France”; French_n_2 “the Romance language spoken in France” and French_a_1 “of or pertaining to France or the people of France”. In Japanese, the first two are productive multiword expressions *furusujiN* “France person” and *furusu-go* “France language” and the third is made by adding the postposition *no* “of” to either of these or just to France: *furusujiN-no* “French (person) lit: France person of”, *furusu-go-no* “French (language) lit: France language of” and *furusu-no* “French (other) lit: France of”. Because these postpositional phrases are completely compositional, it seems redundant to list them in the Japanese Wordnet. In addition, to align accurately, we would have to either separate the current adjective synset into three senses: “of or pertaining to the language of France”; “of or pertaining to the people of France” and “of or pertaining to the France” possibly with the third as the hypernym of the first

²In addition there are two more which are not relevant to this discussion.

two. A better approach may be to take advantage of the rich structure of the current wordnet and allow alignment between *furusu* “France” and *French* through the pertainym relation (*French_a_1* pertains-to *France_n_1*). However, currently there is no easy way to link *furusu-go* “French_n_1 (Language)” with *French_a_1*. Perhaps the proper solution is to add additional pertainym links: *French_a_1* pertains-to *French_n_1* (language) and *French_a_1* pertains-to *French_n_2* (people). Note that similar differences exist, of course, between English and Italian, but they occur far less often due to greater similarity between the two languages.

We have a rich source of new senses suggested by the translators (13,857 cases) that can be used to extend the cover of the Japanese Wordnet. For example, in Figure 2, *remove* is translated as 離す *hanasu*, even though this word was not one of the synonyms for that synset in the Japanese Wordnet. A preliminary investigation of these found that, in all cases, something had to be added to the wordnet, and in 60% of the cases the suggested translation could be used as is. The remaining cases fall into three groups (similar to those discussed above): loose translations which do not really refer to the same synset; Japanese tokens which should be part of a larger multiword expression; and translations which change the part of speech. In addition, we found some errors in the English sense tagging.

In future work, we intend to investigate techniques for efficiently correcting any remaining errors in the corpus. As much as possible, we would like to fix errors in both English and Japanese, so that we can start to carry out quantitative contrastive semantic analysis.

We would also like to investigate how ambiguities are distributed across different languages. For example, 齒 *ha* “tooth” is used for human teeth and cogwheel teeth in English, Japanese and Italian: all three languages share the same ambiguity. In general, we expect to find less ambiguity shared

```

<?xml version="1.0" encoding="utf8"?>
<KAF lang="jpn">
  <kafHeader>
    <fileDesc filename="br-k01"/>
    <linguisticProcessors layer="text">
      <lpt timestamp="2011-09-23T11:45:18" version="0.98" name="MeCab"/>
    </linguisticProcessors>
  </kafHeader>
  <text>
    <wf wid="w1.1.1" sent="1" para="1">スコッティ</wf>
    <wf wid="w1.1.2" sent="1" para="1">は</wf>
    <wf wid="w1.1.3" sent="1" para="1">学校</wf>
    <wf wid="w1.1.4" sent="1" para="1">に</wf>
    <wf wid="w1.1.5" sent="1" para="1">戻ら</wf>
    <wf wid="w1.1.6" sent="1" para="1">なかっ</wf>
    <wf wid="w1.1.7" sent="1" para="1">た</wf>
    <wf wid="w1.1.8" sent="1" para="1">。</wf>
  </text>
  <terms>
    <term tid="t1.1.1" lemma="スコッティ" type="open" pos="N.名詞.一般">
      <span>
        <target id="w1.1.1"/>
      </span>
      <component lemma="スコッティ" id="c1.1.1" pos="N.名詞.一般"/>
    </term>
    <term tid="t1.1.3" lemma="学校" type="open" pos="N.n">
      <span>
        <target id="w1.1.3"/>
      </span>
      <component lemma="学校" id="c1.1.3" pos="N.名詞.一般"/>
      <externalReferences>
        <externalRef resource="Wordnet jpn 1.1" reference="jpn-11-学校-n"/>
      </externalReferences>
    </term>
    <term tid="t1.1.5" lemma="戻る" type="open" pos="V.v">
      <span>
        <target id="w1.1.5"/>
      </span>
      <component lemma="戻る" id="c1.1.5" pos="V.動詞.自立"/>
      <externalReferences>
        <externalRef resource="Wordnet jpn 1.1" reference="jpn-11-戻る-v"/>
      </externalReferences>
    </term>
  </terms>
</KAF>

```

Figure 3: Sample KAF record for スコッティ は 学校 に 戻ら なかつた。 Scotty ha gakkō ni modora nakat ta ., the Japanese translation of English sentence Scotty did not go back to school

between languages from very different families (such as Japanese and English/Italian), but there is also extensive borrowing between English and Japanese. With a sense-tagged tritext, we can start to investigate these questions.

5 Conclusion

In this paper we described the creation of the Japanese Semantic Corpus JSEMCor. The corpus is a translation of the English SEMCOR, with senses projected across from English. The final corpus consists of 14,169 sentences with 150,555 content words of which 58,265 are sense tagged.

References

- Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering*, 11(3):247–261, sep. Special Issue on Parallel Texts.
- Wouter Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. 2009. KAF: a generic semantic annotation format. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL 2009)*, Pisa.
- Jordi Daude, Lluis Padro, and German Rigau. 2003. Validation and tuning of Wordnet mapping techniques. In *Proceedings of the International Conference on Generative Approaches to the Lexicon (GL 2003)*, Pisa.

ence on Recent Advances in Natural Language Processing (RANLP'03), Borovets, Bulgaria.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

W. Nelson Francis and Henry Kucera, 1979. *BROWN CORPUS MANUAL*. Brown University, Rhode Island, 3 edition. (<http://khnt.aksis.uib.no/icame/manuals/brown/>).

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 230–237, Barcelona, Spain, July. Association for Computational Linguistics.

Shari Landes, Claudia Leacock, and Christiane Fellbaum. 1998. Building semantic concordances. In Fellbaum (Fellbaum, 1998), chapter 8, pages 199–216.

A Survey of WordNets and their Licenses

Francis Bond

Linguistics and Multilingual Studies
Nanyang Technological University
bond@ieee.org

Kyonghee Paik

Waseda Media Network Center and
Center for Modern Languages, NTU
paikbond@gmail.com

Abstract

This paper surveys currently available wordnets. We measure the effect that license choice has on their usage, measured by the number of citations. Finally, we discuss methods to make wordnets more generally accessible, starting with a shared online server for freely distributable wordnets.

1 Introduction

In this paper we answer two questions: (i) what effect does license choice have on wordnet uptake? and (ii) How can we make wordnets more widely useful? To answer these questions we start off by surveying the current available wordnets and end up trying to make a multilingual wordnet server.

This paper was mainly inspired by two things. The first was a discussion with researchers in Europe who were presenting their work on linking semantic classes between French and Italian. They were using semantic classes derived from clustering Europarl and we asked why they didn't use WordNet. Their answer was that it wasn't available for their languages. Our first reaction was shock — of course there are wordnets for French and Italian. On relection, we realized that in fact, they are somewhat hard to find. If you search for downloadable wordnets for these two languages, you end up at the ELRA page, where you are charged a considerable amount of money for out-of-date versions from the EuroWordNet project. If you know what to look for, you can find free (at least for research) up-to-date versions of wordnets for French and Italian, called WOLF and MultiWordNet respectively, but they are not obvious.

The second inspiration was discussions within projects we have been involved with

building wordnets, where we were trying to decide which license should be used. We found there was very little in the way of qualitative evidence that one license was better than another, and decided to try to produce some ourselves. There are over 40 projects to build wordnets for various languages. They range from the Princeton WordNet of English (Fellbaum, 1998) the original wordnet project which has over 150,000 concepts, to research projects such as those on Bantu or Norwegian, which have yet to release any results. Because there are so many wordnets all sharing a similar structure, but with a wide variety of licenses, they provide good data to look at license use.

Language resources, to be useful, must be both **accessible** (legally OK to use) and **usable** (of sufficient quality, size and with a documented interface) (Ishida, 2006). We address both of these concerns in this paper.

This paper is structured as follows. First, we survey currently available wordnets (§ 2). Based on the results we look at the effect that license choice has on wordnet citations (§ 3). Then we look at building a combined multilingual wordnet based on the available free wordnets (§ 4). We finish with some discussion (§ 5) and then conclude.

2 A Survey of WordNets and their Licenses

We have compiled a survey of wordnet projects, and found the license for the projects that have released data (see Table 1). This table shows the project name, languages included, number of synsets, first release, license, canonical citation and number of citations for the canonical citation.

Roughly a third of the wordnets are **open source**, that is, free and redistributable with

no constraints. The most common license is a variant of original Princeton Wordnet License (a modified MIT license), the rest are free licenses such as the LGPL, GNU-FDL, GPL or CC BY. These are all free, open-source licenses according to widely accepted definitions such as the Open Source Definition,¹ Debian Free Software Guidelines² or the GNU project.³

Roughly a third are available **free for research**, but they cannot be redistributed or have some restriction on their use and are thus not open source licenses. Most of these also offer a separate license for commercial use.

Finally, there are roughly a third of the wordnets which are **non-free**, costing money even for research use, although generally with a reduced cost. Some wordnets have both free and non-free versions. In particular, wordnets produced through Euro WordNet are all sold by the European Language Resources Association (ELRA) even if exactly the same data or updated newer versions are also available for free.

There is a trend for newer projects to use open licenses (such as Japanese, Finnish and Thai) and for older projects to re-release their data under more open licenses (such as German and Catalan).

To compile this data we took as our starting point the table of WordNet projects *Wordnets in the World*⁴ maintained by the Global WordNet Association (GWA). In addition to projects listed here, we have added other projects that we discovered through mailing lists or conferences.

As our interest is mainly in NLP applications, we have not included wordnets which have an online interface but no information on how to obtain the whole database (such as Latin, Nepali, Portuguese, and many others). We have done our best to survey as many wordnets as we can, but apologize in advance if we have missed any. There are some projects (e.g., Albanian) where we could actually download the data, but could not find information on the license, we have omitted them from the table.

¹www.opensource.org/docs/definition.php

²www.debian.org/social_contract.html

html#guidelines

³www.gnu.org/licenses/license-list.html

⁴www.globalwordnet.org/gwa/wordnet_table.htm

Size	Date	Open	Free	Non free
Large	2008	Japanese 24	Dutch 19	
Large	2009	Danish/Thai 8/4		Korean 5
Small	2008	French 22	Slovenian 13	Bulgarian 3

Table 2: Similar wordnets with different licenses

We sent out a questionnaire to get more data for the paper. Entries in the table marked with a * are based on the questionnaire responses. The questionnaire is given in Appendix A.

We have created a map based on this data (Figure 1). The map is based on countries, which means that language/mapping involves some subjective judgment.⁵ Languages with an open source wordnet are shown in green, with a free for research wordnet in blue, and a non-free wordnet in brown. The higher the number of synsets, the lighter the color. When there are two wordnets with different licenses, we have used the most complete version as representative for that languages.

The map shows that much of the world has at least some wordnet for it, although the coverage of Africa and central Asia is still very incomplete.

3 Assessing the Effect of License Choice on WordNet Citations

In order to compare effect of license choice on number of citations, we compare a few similar size projects released at similar times in Table 2. Any comparisons done in this way are inherently noisy — Google Scholar counts may be wrong and citations may depend on any number of factors, including the novelty of the construction method, the number of people working on the language, the accessibility of the publication venue and so forth. However, comparing the most similar pairs we can, in general the more free version has the most citations.

There are some exceptions, the Chinese wordnet (Xu et al., 2008) has no citations, despite it being large and **free for research**. We suspect that this is because you must email

⁵For example, we have marked Spain as having free coverage thanks to the Catalan Wordnet.

Name	Language	# Synsets	Release	License	Citation	Count
Open Source						
Princeton WN* [€]	English	155,000	1991	WordNet	Fellbaum (1998)	6,821
FinnWordNet	Finnish	117,700	2010	WordNet	Lindén and Carlson. (2010)	0
Russian WN	Russian	117,000	2004	Wordnet	Balkova et al. (2008)	15
Thai Wordnet	Thai	73,593	2007	WordNet	Thoongsup et al. (2009)	4
DanNet*	Danish	65,000	2008	WordNet	Pedersen et al. (2009)	8
Japanese WN*	Japanese	57,000	2009	WordNet	Isahara et al. (2008)	24
Catalan WN*	Catalan	42,000	1999	GPL	Benítez et al. (1998)	17
LSG	Irish Gaelic	32,742	?	GNU FDL	http://borel.slu.edu/lsg/	—
Hindi WN	Hindi	28,687	?	GNU FDL	Jha et al. (2001)	10
WOLF	French	22,000	2009	Cecill-C [†]	Sagot and Fišer (2008)	22
Wordnet Bahasa*	Malay, Indonesian	20,000	2011	MIT	Nurrial Hirfana et al. (2011)	—
Spanish WN* [○] [€]	Spanish	15,556	2006	LGPL	Farreres et al. (1998)	65
Catalan WN* [○] [€]	Catalan	15,556	2006	LGPL	Benítez et al. (1998)	17
Arabic WN*	Arabic	11,269	2008	CC BY SA	Black et al. (2006)	28
Hebrew WN*	Hebrew	5000	2006	GPL	Ordan and Wintner (2007)	0
Free for Research						
Chinese WN*	Chinese	115,424	2008	res/com	Xu et al. (2008)	0
KorLex* [○]	Korean	90,000	2007	res/com	Yoon et al. (2009) (nouns)	—
Spanish WN* [€]	Spanish	62,000	1999	res/com	Farreres et al. (1998)	65
Cornetto* [€]	Dutch	70,371	2009	res/com	Vossen et al. (2008)	19
GermaNet* [€]	German	69,594	2011	res/com	Kunze and Lemnitzer (2002)	52
MultiWN* [€]	Italian	38,877	2008	res/com	Pianta et al. (2002)	143
MWN*	Macedonian	33,276	2010	CC BY NC	Saveski and Trajkovski (2010)	0
Ro-WordNet*	Romanian	30,000	soon	no-deriv.	Tufiš et al. (2008)	9
Czech WN * [€]	Czech	29,000	1999	res/com	Pala and Smrž (2004)	34
SloWnet*	Slovene	20,000	2010	CC BY NC SA	Fišer and Sagot (2008)	13
Non Free (Available for Research)						
KorLex*	Korean	130,878	2007	res/com	Yoon et al. (2009)	5
Estonian * [€]	Estonian	47,000	—	ELRA	Kerner et al. (2010)	0
EuroWordNet					Vossen (1998)	728
Dutch	Dutch	44015	1999	ELRA	ELRA-M0016	
Spanish	Spanish	23370	1999	ELRA	ELRA-M0017	
Italian	Italian	48529	1999	ELRA	ELRA-M0018	
German	German	15,132	1999	ELRA	ELRA-M0019	
French	French	22,745	1999	ELRA	ELRA-M0020	
Czech	Czech	22,745	1999	ELRA	ELRA-M0021	
Estonian	Estonian	9,317	1999	ELRA	ELRA-M0022	
ItalWordNet	Italian	49,360	1999	ELRA	ELRA-M0018	
BasqWN	Basque	30,281	?	ELRA	Pociello et al. (2011)	0
BulNet* [○]	Bulgarian	23,715	2004	ELRA	ELRA-M0041 (Koeva, 2008)	3

Table 1: Catalog of WordNets

^{*}Results from our survey[○]A subset released under a less restrictive license[€]A version from EuroWordNet is also available from ELRA[†]A variant of the LGPL

res/com means that it is available under different license for research and commercial use

Release is the first release under this license

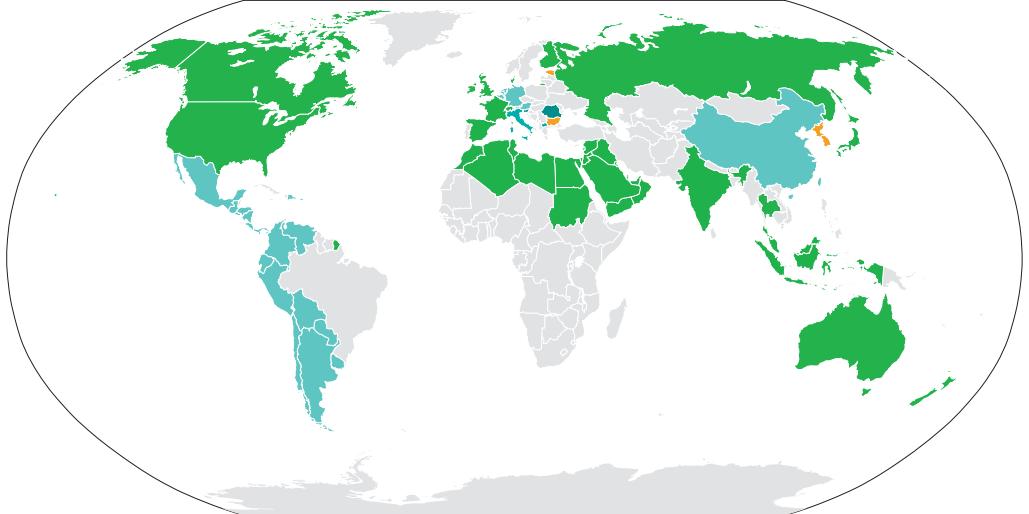


Figure 1: Map of Countries showing WordNet availability

Countries with **open source wordnets** are in green; **free for research** wordnets are in blue; **non free** wordnets are in brown. The lighter the color, the more synsets.

Citation counts from Google Scholar (accessed on 2011-09-23)

and ask for permission to use it, which is a substantial barrier to use. The Italian wordnet (Pianta et al., 2002) has a very high number of citations. In this case it was developed as part of a multilingual wordnet with several other languages, thus giving it a large citation group. Finally, the Thai wordnet (Thoongsup et al., 2009), has relatively few citations, in this case it is also a part of a large project (the Asian Wordnet: Sornlertlamvanich et al. (2008)) which gets more citations (10).

Even with all these caveats, we think that the data supports the unsurprising result that the more open the license a wordnet is released under, the more likely it is to be used (or at least cited). In other word, uptake of a resource depends on how **usable** (legally accessible) a resource is.

4 Construction of an Open Multi-lingual WordNet

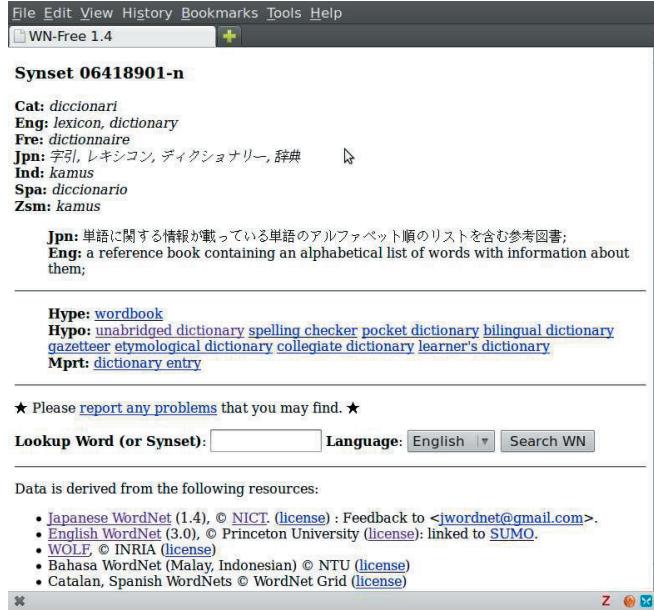
In order to make the wordnets more **accessible**, we have started to build a simple SQL server with information from those wordnets whose licenses allows us to do so. We show a screen shot in Figure 2. Most of these wordnets are based on the **extend** model, basically

adding lemmas in new languages to existing English synsets. Therefore, adding a new language is a case of just adding new lemmas to the synsets (annotated with their language). In theory, this should be easy.

In practice, adding a new language turned out to be difficult for two reasons. The first problem was that none of the wordnets we surveyed updated their structure when the English wordnet did. In order to combine them into a single multilingual structure, we had to map to a common version. The second problem was the incredible variety of formats that the wordnets are distributed in. Almost every project used a different format and thus required a new script to convert it. These two problems mean that, even if a wordnet is legally available, there is still a technical hurdle before it becomes easily accessible.

The first problem can largely be overcome using the mapping scripts from Daude et al. (2003). Mapping introduces some distortions, in particular, when a synset is split, we chose to only map the translations to the most probable mapping, so some new synsets will have no translations.

The second problem we are currently solving



Catalan, English, French, Japanese, Indonesian, Spanish and Malaysian wordnets stored in a common database searchable database.

Figure 2: Screenshot of Free Wordnet Lookup

through brute force, writing a new script for almost every new wordnet we add. We discuss better possible solutions in Section 5.

The server currently includes English (Fellbaum, 1998), Japanese (Bond et al., 2008); French (Sagot and Fišer, 2008); Indonesian and Malaysian (Nurrial Hirfana et al. 2010); Catalan and Spanish from the Global WordNet Grid (Fig 2.) The wordnets are all in a shared `sqlite` database with a PERL cgi server using the wordnet module produced by the Japanese WordNet project (Isahara et al., 2008).

We plan to add further free wordnets. Additional wordnets with licenses that allow us to serve the data exist for Arabic, Danish, Finnish, Gaelic, Hebrew, Hindi, Macedonian, Slovenian and Thai (see Table 1 or 3).

We also have a script that outputs wordnets from our database into either tab separated values, where they can be used by the Natural Language Tool Kit⁶ (Bird et al., 2009) or the emerging standard of WordNet-LMF (Lex-

⁶With the extensions that were added with the Japanese translation (Bird et al., 2010).

ical Markup Framework: Soria et al., 2009). Finally, we will also make the SQL database available. Licenses that allow redistribution of derivative works allow people to make the entire lexicons available in any format, thus greatly improving their usefulness.

Notes on the various formats and conversions

Our conversion scripts basically reduce each wordnet to a list of synset-lemma pairs, which we then map to the English 3.0 synsets. We currently lose any extra information about, for example, other morphological forms of words (we expect the morphological analyzers to give us lemmas). We also lose any synsets not in the English 3.0 wordnet.

T/CSV (Tab/Comma Separated Values) The Japanese Wordnet is already available as a table of synset-lemma pairs, and a third field that gives a confidence score, so we can use it immediately. The Bahasa Wordnet has an additional field that says whether each synset-lemma pair can be used in Malay or Indonesian or either (98% of entries are ac-

Wordnet	Ver	Format
Arabic	?	pwn
Bahasa✓	3.0	tsv
Catalan✓	1.6	gwa-xml
Danish	2.0	csv, owl
English	3.0	pwn*
Finnish	3.0	text tables
French (WOLF)✓	2.0	xml
Irish		
Japanese✓	3.0	tsv
Hebrew		xml
Hindi	?	pwn
Spanish✓	1.6	gwa-xml
Thai	3.0	LMF

Table 3: Free WordNets and their formats

*Read using the perl WordNet::QueryData module
(Rennie, 2000)

✓ Conversion script built

Ver. is the corresponding English version

ceptable in either language). We pre-process the wordnet to output two wordnets, one for Malay and one for Indonesian. The Danish wordnet is similar to the Japanese wordnet.

GWA-XML A subset of the Spanish and Catalan wordnets are released under the LGPL. They are released in pseudo XML. We convert them to actual XML by fixing the encoding of the quotation marks and adding a tag surrounding the whole file (`<wn>...</wn>`⁷). French (WOLF) and Hebrew are released in a very similar format. This format is based on the de facto standard established by the Euro WordNet project.

We have a simple script that pulls out the synset-lemma pairs from this XML (and ignores everything else).

LMF To get a mapping from a lemma to the English synset in LMF theoretically requires two mappings (lemma to language synset, language synset to English synset). In practice, for the current version of the Thai Wordnet, the Thai synset ID is always the same as the English 3.0 ID. We take advantage of this to extract the information we want with a very simple script.

⁷We have reported these problems upstream.

The remainder of the wordnets all use quite different formats. We are hoping to persuade each individual project to also output the data in a more universal format.

5 Discussion and Further Work

In general, the freer the license, the more a wordnet appears to be used. Therefore, for projects whose goal is to produce a resource that will be widely used, the freer the license the better.

Access to multiple wordnets would, of course, be simplest if everyone used the same format. This would also make tool sharing easier and perhaps reduce some of the current duplicated effort. Currently there seem to be two front runners: Wordnet-LMF (Soria et al., 2009), which is explicitly designed as an interchange format, and Wordnet SQL.⁸ Both of them are easily processed with existing interfaces (for XML and SQL) unlike the PWN format, which is very specialized to wordnet. However, different projects already have invested effort into their current interfaces and formats, so may not change quickly. In addition, it is often easier to get funding to build resources that to maintain them.

Given this, a more realistic medium term goal for increasing access to multiple wordnets is to encourage conversion from whatever local representation to a shared representation, such as Wordnet-LMF. Because accurate conversion relies on knowledge of each wordnet’s individual structure, it would be safer if each project did this conversion themselves. As a start, we will release our conversion scripts.

In the near future, we also plan to improve our conversion scripts so as to also add new synsets when they exist, although there is currently no way to link them across different language pairs. This problem was solved with the Inter-Lingual Index in EuroWordNet (Vossen, 1998), but currently there is not yet a single index shared by all projects.

All of the data in our catalog, and the map, is being fed back to the *Wordnets in the World* page. We have already been sending additions and corrections to the maintainers throughout the project. In particular, we found the contact details were out of date for 5 out of 45

⁸wnsql.sourceforge.net/uml2.html

projects, and are in the process of finding the current contacts for these projects. We are also adding projects that we know of through the Asian WordNet meetings to the list.

Finally, on a positive note, regardless of the actual license, researchers are generally very willing to share their data, and will often make it available on request, or even link to it online, even though the license does not, strictly speaking allow this. While this is very welcome, receiving data without a proper license does not legally allow its use, and thus does not lead to more reproducible research.

6 Conclusions

We have surveyed the current coverage of wordnets, both in terms of size and license. Many of the world's most widely used languages now have wordnets, although not all of them are freely available, and lack of standard interfaces and data formats makes them hard to access. We have made a first step to increasing accessibility by converting free wordnets to a common format. We show that, in general, wordnets released with freer licenses are cited more often.

Acknowledgments

We would like to thank the compilers of the Global WordNet Association's *Wordnets in the World* page as well as everyone who responded to the survey. This research was funded by the Creative Commons Catalyst Grant: *Assessing the effect of license choice on the use of lexical resources*.

References

- Valentina Balkova, Andrey Sukhonogov, and Sergey Yablonsky. 2008. Russian wordnet: From UML notation to internet/intranet database implementation. In Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Second International WordNet Conference — GWC 2004*, pages 31–38. Brno.
- L. Benítez, S. Cervell, G. Escudero, M. López, G. Rigau, and M. Taulé. 1998. Methods and tools for building the Catalan wordnet. In *ELRA Workshop on Language Resources for European Minority Languages*. Granada, Spain.
- Stephen Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly. (www.nltk.org/book).
- Stephen Bird, Ewan Klein, and Edward Loper. 2010. *Nyuumon Shizen Gengo Shori [Introduction to Natural Language Processing]*. O'Reilly. (translated by Hagiwara, Nakamura and Mizuno).
- W. Black, S. Elkateb, H. Rodriguez, M. Alkhalfa, P. Vossen, A. Pease, M. Bertran, and C. Fellbaum. 2006. The Arabic wordnet project. In *Proceedings of LREC 2006*.
- Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008. Boot-strapping a WordNet using multiple existing WordNets. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.
- Jordi Daude, Lluís Padro, and German Rigau. 2003. Validation and tuning of Wordnet mapping techniques. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'03)*. Borovets, Bulgaria.
- X. Farreres, G. Rigau, and H. Rodríguez. 1998. Using wordnet for building wordnets. In *COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems*. Montreal, Canada.
- Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- Darja Fišer and Benoît Sagot. 2008. Combining multiple resources to build reliable wordnets. *Text, Speech and Dialogue*, LNCS 2546:61–68.
- Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.
- Toru Ishida. 2006. Language grid: An infrastructure for intercultural collaboration. In *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, pages 96–100. URL <http://langgrid.nict.go.jp/file/langgrid20060211.pdf>, (keynote address).
- S. Jha, D. Narayan, P. Pande, and P. Bhattacharyya. 2001. A wordnet for Hindi. In *International Workshop on Lexical Resources in Natural Language Processing*. Hyderabad.
- Kadri Kerner, Heili Orav, and Sirli Parm. 2010. Growth and revision of Estonian wordnet. In Pushpak Bhattacharyya, Christiane Fellbaum, and Pieck Vossen, editors, *5th Global Wordnet Conference: GWC-2010*, pages 198–202. Mumbai.
- Sv Koeva. 2008. Derivational and morphosemantic relations in Bulgarian wordnet. In *Intelligent Information Systems, XVI*, pages 359–389. Academic Publishing House, Warsaw.
- C. Kunze and L. Lemnitzer. 2002. Germanet — representation, visualization, application. In *LREC*, pages 1485–1491.
- Krister Lindén and Lauri Carlson. 2010. Finnwordnet — wordnet på finska via översättning. *LexicoNordica — Nordic Journal of Lexicography*, 17:119–140. In Swedish with an English abstract.
- Nurril Hirfana Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open Wordnet baha. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*. Singapore. (to appear).
- Noam Ordan and Shuly Wintner. 2007. Hebrew wordnet: a test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1):39–58.
- Karel Pala and Pavel Smrž. 2004. Building Czech wordnet. *Romanian Journal of Information Science*, 7:79–88.

- B.S Pedersen, S. Nimb, J. Asmussen, N. Sørensen, L. Trap-Jensen, and H. Lorentzen. 2009. DanNet — the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*.
- Emanuele Pianta, Luisa Bentivogli, and Christian Giardi. 2002. Multiwordnet: Developing an aligned multilingual database. In *In Proceedings of the First International Conference on Global WordNet*, pages 293–302. Mysore, India.
- Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and construction of the Basque wordnet. *Language Resources and Evaluation*, 45(2):121–142.
- Jason Rennie. 2000. Wordnet::querydata: a Perl module for accessing the WordNet database. <http://www.ai.mit.edu/people/jrennie/WordNet>.
- Benoit Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.
- M. Saveski and I Trajkovski. 2010. Automatic construction of wordnets by using machine translation and language modeling. In Jernej Zganec Gros, Ljubljana Tomaz Erjavec, editor, *In Proceedings of Seventh Language Technologies Conference, 13th International Multiconference Information Society*, volume C.
- Claudia Soria, Monica Monachini, and Pieck Vossen. 2009. Wordnet-LMF: fleshing out a standardized format for wordnet interoperability. In *Second International Workshop on Intercultural Collaboration (IWIC-2009)*. Stanford.
- Virach Sornlertlamvanich, Thatsanee Charoenporn, Kergit Robkop, and Hitoshi Isahara. 2008. KUI: Self-organizing multi-lingual wordnet construction tool. In Attila Tanács, Dóra Csédes, Veronika Vincze, Christiane Fellbaum, and Pieck Vossen, editors, *4th Global Wordnet Conference: GWC-2008*, pages 417–427. Szeged, Hungary.
- Sareewan Thoongsup, Thatsanee Charoenporn, Kergit Robkop, Tan Sinthurahat, Chumpol Mokarat, Virach Sornlertlamvanich, and Hitoshi Isahara. 2009. Thai wordnet construction. In *Proceedings of The 7th Workshop on Asian Language Resources (ALR7), Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP)*, Suntec, Singapore.
- Dan Tufiș, Radu Ion, Luigi Bozianu, Alexandru Ceaușu, and Dan Ștefănescu. 2008. Romanian wordnet: Current state, new applications and prospects. In *Proceedings of the 4th Global WordNet Association Conference*, pages 441–452. Szeged.
- P. Vossen, I. Maks, R. Segers, and H Van der Vliet. 2008. Integrating lexical units, synsets and ontology in the Cornetto database. In *LREC 2008*. European Language Resources Association (ELRA), Marrakech, Morocco.
- Piek Vossen, editor. 1998. *Euro WordNet*. Kluwer.
- Renjie Xu, Zhiqiang Gao, Yuzhong Qu, and Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual Chinese-English WordNet. In *3rd Asian Semantic Web Conference (ASWC 2008)*, pages 302–341.
- Aesun Yoon, Soonhee Hwang, Eunrourng Lee, and Hyuk-Chul Kwon. 2009. Construction of Korean wordnet KorLex 1.5. *Journal of KIISE: Software and Applications*, 36(1):92–108.

A Questionnaire about WordNet licensing

This is the questionnaire we sent out, reformatted to fit in one column. We also added a filled-out sample.

Please send to: xxx@yyy

More information can be found at:
<http://zzz.yyy>

Questions:

* Resource Name:

* Language(s) Described:

* Developer(s):
 (Institution)

* Contact Person(s):
 (name and email please)

* URL:

* Date of (planned) release:

* License:

* Reason for choosing license:

* Current Size (in synsets):

* Canonical citation(s):

* Funding Source(s):

* Main users:

Please answer the questions and send the questionnaire to <xxx@yyy>. If you have any other questions or comments, feel free to ask us.

If you would like us to not add the data to the GWA page, please let us know. Otherwise we will do so.

We estimate that filling in this form should take around 15-20 minutes.

Semantic Hand-Tagging of the SenSem Corpus Using Spanish WordNet Senses

Irene Castellón

GRIAL, UB

Barcelona, Spain

icastellon@ub.edu

Salvador Climent

GRIAL, UOC

Barcelona, Spain

scliment@uoc.edu

Marta Coll-Florit

GRIAL, UOC

Barcelona, Spain

mcoll@uoc.edu

Marina Lloberes

GRIAL, UOC

Barcelona, Spain

mlloberes@uoc.edu

German Rigau

IXA, EHU/UPV

Donostia, Spain

german.rigau@ehu.es

Abstract

This paper presents the semantic annotation of the SenSem Spanish corpus, a research focused on the semantic annotation of the nominal heads of the verbal arguments, with the final goal of acquiring semantic preferences for verb senses. We used Spanish WordNet 1.6 senses in the annotation process. This process involves the analysis of the adequacy of WordNet for semantic annotation and, in cases of inadequacy, the proposal of solutions. The results are the tagged corpus, a guide with semantic annotation criteria, and a critical assessment of WordNet as a resource for semantic corpus tagging.

1 Introduction

The semantic annotation of corpora provides a basis for characterizing lexical-semantic information. In this paper we present a relevant part of the long-term project of annotation of the SenSem Spanish corpus (Alonso et al., 2007), i.e. the semantic annotation of the nominal heads of the verbal arguments using the Spanish WordNet 1.6 (hereinafter ESPWN1.6) (Atserias et al., 2004a). The final goal of this research is the acquisition of semantic preferences for verb senses.

This process involved two lead-up tasks:

- 1) Assessing the adequacy of ESPWN1.6 (and thus WordNets in general) for corpus annotation.
- 2) Finding solutions for such inadequacy problems — thus setting an annotator's guide providing stable criteria, specially for sense disambiguation.

SenSem is a databank of Spanish which maps a corpus and a verbal database. The corpus consists of 25,000 sentences, 100 for each of the 250

most frequent verbs of Spanish (Davies, 2002). Sentences are tagged at both syntactic and semantic levels: verb sense, phrase and construction types, aspect, argument functions and semantic roles.

Currently, the project finished the semantic tagging of the heads of the verbal arguments. This phase of the project consists of tagging noun heads with ESPWN1.6, — a resource integrated into the Multilingual Central Repository¹ (MCR) (Atserias et al., 2004b) which follows the EuroWordNet architecture (Vossen, 1998) and currently maps multiple wordnets and ontologies, e.g. Top Ontology (Álvez et al., 2008) and SUMO (Niles and Pease, 2003).

The next section comments on the state of the art in semantic annotation of corpus and §3 introduces our methodology. Then §4 presents the assessment of ESPWN1.6 as a resource for semantic tagging; and §5 shows the set of solutions and annotation criteria stated to solve problems arisen. The results of the project are presented in section 6 and, finally, conclusions and proposed future work conclude the paper.

2 Semantic Annotation of Corpora

There are several semantically-tagged corpora for English such as PropBank (Palmer et al., 2003) or FrameNet (Baker et al., 1997). However, the projects that are more closely related to ours are Ontonotes (Yu et al., 2007), SemCor (Miller et al., 1994) and Multi-SemCor (Bentivogli and Pianta, 2005) as they use WordNets for sense tagging.

For Spanish, two main semantically-annotated corpora have been developed: ADESSE (García Miguel and Albertuz, 2005) and AnCora (Taulé

¹<http://adimen.si.ehu.es/web/MCR>

et al., 2008). Both consist of a corpus annotated with references to a verbal database.

SenSem is structured in the same way, being its main difference that (i) it is not *a posteriori* mapping since from the beginning it was designed and built in order to annotate corpus data with verb sense information from the database; (ii) it is a balanced representation as each verb lemma holds 100 sentence examples; and (iii) it applies to the most frequent verbs of the language.

3 SenSem Methodology

Our methodology draws upon that of Eusemcor (Agirre et al., 2006) which carried out a concurrent processes of corpus annotation, correction and extension of WordNet for the Basque language. However, for several practical reasons, we did not modify the Spanish WordNet. We just keep record of the potential changes to be incorporated when necessary.

The process presented here was performed by six linguists, three annotators and three annotators and judges. The task spent 6 months and its cost was 8 person/month. The process was divided into the following steps:

- 1) Automatic POS tagging of the corpus using FreeLing (Padró et al., 2010). This process led to identifying the heads of the nominal arguments.
- 2) Adapting Eusemcor interface to the specific needs of the project.
- 3) A preliminary test in order to set an agreement between annotators for developing the initial criteria.
- 4) Full corpus annotation.
- 5) In parallel with 4, coordination sessions with judges and annotators for establishing, extending or modifying the initial criteria.

The annotation was performed using an interface implemented as a web service which facilitates the annotation of all occurrences of a single lemma in the corpus². This ensures annotation consistency as the same linguist deals with all instances of the same word and, in turn, it allows him/her to get a holistic view of the distribution of the lemma in different ESPWN1.6 meanings.

The annotation of the test phase was performed by a team of four linguists (Carrera et al.,

2008). They all annotated the same subset of the corpus with a total of 50 sentences. Instructions to annotators for such preliminary test were quite general — they are listed in §5.1 as *general instructions*.

The initial agreement between annotators was only 40%. Then, we discussed the annotation decisions for producing a first set of disambiguation criteria. A subsequent annotation of a subset of the corpus based on this consensus reached an agreement between annotators of 84.32%. The establishment of specific criteria for disambiguation and semantic annotation continued until they became stable.

This process led to a general assessment of the drawbacks of the nominal part of ESPWN1.6 as a resource for semantic annotation which can be quite straightforwardly ported to other wordnets in similar projects for other languages. This assessment is presented in the following section.

4 Assessment of ESPWN1.6 as a Resource for Semantic Tagging

We have classified the drawbacks found when using ESPWN1.6 for semantic annotation into five general types: technical (§4.1), structural (§4.2), derived from ambiguity or uncertainty of meaning (§4.3), related to incompleteness (§4.4), and interlinguistic (§4.5).

4.1 Technical problems

Firstly, let us consider some problems derived from decisions taken on the original design of WordNet and ESPWN1.6.

Lexicographical description. WordNet glosses and/or examples might conflict with the meaning one can infer from the ontological relations of the concept.

Morphology. ESPWN1.6 was derived by mapping and translating the English WordNet version 1.6. Thus, it encodes as lemmas English feminine forms which in Spanish are inflectional (e.g. *sister*-‘hermano/a’). Similar problems are found with diminutives. This causes mismatching between FreeLing and ESPWN1.6 lemmatization criteria thus leading to several problems in annotation.

4.2 WordNet 1.6 structural drawbacks

Since ESPWN1.6 follows the expand approach (Vossen 1998) it also inherits the English WordNet structural problems.

²<http://ixa2.si.ehu.es/spsemcor/>

Autohyponymy. Quite often two senses of the same word exhibit a direct hyponymy relationship. In these cases, WordNet encodes different levels of generalization of the same concept.

For instance, the lemma ‘trabajo’ (*work*) in ESPWN1.6 spans in seven senses, including ‘trabajo_4’ (“productive work”) and ‘trabajo_1’ (“activity directed toward making or doing something”); this two synsets are related by hyponymy —‘trabajo_4’ ISA ‘trabajo_1’.

False hyponymy. The taxonomic structure of WordNet encodes ontological inaccuracies. Guarino (1998) classifies such false hyponymy into the following categories:

- Confusion of meaning in situations of multiple inheritance. Incompatible meanings collapse into one synset, e.g. ‘Statue of Liberty’ having both hypernyms ‘statue’ (an object) and ‘plastic_art’ (an abstract concept).
- Reduction of meaning e.g. ‘organization’ as hyponym of ‘group’ — while the meaning of *organization* is wider than that of *group*, not an specialization of it.
- Overgeneralization. It happens in case of excessive heterogeneity of cohyponyms so that indirectly WordNet forces an inappropriate more general interpretation of the hyperonym — e.g. *an amount of matter* as hyponym of a *physical object*.
- Confusion between type and role. This is a common mistake in the WordNet taxonomy. A taxonomic relation by definition must be established between entity types but in WordNet there are relations from type to role, e.g. ‘persona_1’ (*person*) as a hyponym of ‘agente_causal_1’ (*causal agent*).
- Relationship confusion. One of the most common structural problems is the confusion between taxonomy and meronymy; e.g. ‘hueso_1’ (*bone*) shows up as hyponym of ‘connective_tissue_1’ when in fact *bone* is an entity “made of” *tissue*, so the relationship should be meronymy.

4.3 Ambiguity and vagueness of meaning

Excessive granularity of WordNet senses is the most mentioned problem in the literature, especially that concerning Word Sense Disambiguation (Agirre and Edmonds 2007). As a direct consequence of this problem semantic distinctions are difficult to be drawn by human annota-

tors: for the same lemma very similar senses can be possible. Moreover, regular distinctions, such as various types of regular polysemy (Apresjan, 1973) are not applied consistently in WordNet.

We can classify the general problem of excessive granularity of meaning in WordNet on the following basic types:

Regular Polysemy. Some entities or situation types are systematically polysemous, as in the case of events and its resulting state. For example, in one occurrence of “the education of youth”, it is difficult to discern whether someone is talking about either the event or the final state of the person as “educated” — both possibilities are synsets in ESPWN1.6.

Sense proliferation because of the point of view. Frequently, many different WordNet synsets actually denote the same type of entity but being observed from different points of view. For instance, ‘familia’ (*family*) has three senses in ESPWN1.6 corresponding respectively to “a social unit living together”, a “primary social group” and “people descended from a common ancestor”.

Senses modulated by context. Only a very rich context could allow annotators to disambiguate between the possible meanings of a word. For instance, ‘interno’ has three meanings in ESPWN1.6: (i) a child in a boarding school, (ii) a person confined in an institution (like a prison or hospital) or (iii) a resident doctor in a hospital. Annotation is performed on a sentence basis so usually there is no context enough to draw such subtle differences.

Sense intersection. In some cases two meanings of a word do not stand for completely disjoint aspects of the denotation — instead, they intersect.

Clearly these types of ambiguity are facts of language, not drawbacks of WordNet. The point here is that WordNet’s developers rather than address these problems in some simple and structured way, they choose to add new word meanings, not always consistently.

A different but related problem arises not from WordNet design but from the information it provides:

Insufficient or misleading information. When neither the gloss nor the examples or the structure of ESPWN1.6 help to make the semantic distinctions clear.

4.4 Problems of ESPWN1.6 incompleteness

WordNet, despite being the largest lexical-semantic knowledge base, does not include all the existing vocabulary of a language. The main dimensions of its incompleteness are the following:

Lack of synset. There are lemmas not appearing in ESPWN1.6. Moreover also the resource does not contain the corresponding interlingual concept, e.g. ‘seguridad social’, whose English translation would be something like *social health system*.

Lack of variant. Some synsets are incomplete as they do not bear some Spanish synonyms. For instance, ‘testamento vital’ is not encoded in ESPWN1.6 but it is in the English WordNet.

Metaphorical senses. WordNet does not include in a systematic way those meanings created by metaphorical extension. For example, ESPWN1.6 ‘puente’ (*bridge*) does incorporate the original architectural sense plus those senses denoting a dental prothesis and the gymnastics exercise, but it does not have the sense of a special bank holiday in Spain.

Moreover, in some cases ESPWN1.6 includes the metaphorical sense but not the original one, as in ‘pie’ (*foot*), which bears the meaning corresponding to the base of a mountain but surprisingly not that of the limb of a person.

These are both examples of incompleteness of the building process of ESPWN1.6 from the English WordNet. In the first case because the meaning of ‘bridge’ as “holiday” is not lexicalized in English and in the second because of a simple oversight.

Lack of multiword units. Similarly, ESPWN1.6 implements multiword units without following consistent criteria, e.g. it includes ‘agente secreto’ (*secret agent*), which has a non-compositional reading, and also ‘police officer’, which has a compositional reading.

Lack of named entities. Proper names in WordNet are mostly culturally related to the United States. ESPWN1.6 incorporates a number of named entities but of course not all of the existing ones.

4.5 Interlinguistic problems: language versions of WN1.6

Several reported problems are caused by the way ESPWN1.6 was built, i.e. as an expansion of the

English WordNet so that the level of adaptation of WordNet into Spanish is very limited. In this section we list the drawbacks straightforwardly caused by this.

Lack of Spanish equivalent. There are cases where the concept does exist in the English WordNet but the ESPWN1.6 has not the equivalent translation. For instance, ‘juez’ (*judge*) is monosemous in ESPWN1.6 as it appears only with the usual sense related to the legal system. However the sense of “evaluator”, which is present in the English WordNet (‘judge_2’, ‘evaluator_1’) is not present in ESPWN1.6.

Synsets split because of morphological mismatches. Since English has no grammatical gender there are cases in which two lemmas of English correspond to only one in Spanish. This causes the inappropriate splitting of the Spanish word in two different synsets, as in ‘tia_1’ and ‘tio_2’ as a result of the projection of English ‘aunt_1’ and ‘uncle_1’.

Cultural bias. Some concepts in ENGWN1.6 are culturally marked. This is improperly projected to Spanish thus causing several imbalances in ESPWN1.6. For example, the lemma ‘presidente’ (*president*) does not include the case of the head of the government in a kingdom or similar type of state; WordNet only has the President as the head of a (republican) state.

5 Solutions and Annotator's Guide

This section details the solutions adopted for the problems outlined above, which have been collected in a guideline for the annotators. Most are pragmatic solutions, often conditioned to the initial requirement of performing the annotation without changing the original ESPWN1.6. This decision relied on the fact that their improvement was out of the scope of the project and we had not the rights to do it. However, the assessment described in §4 is a good base to face the task in future projects.

5.1 Technical problems

General instructions. The following instructions have been defined in order to carry out the annotation process:

- The relations of a synset should be always inspected, at least the hypernym(s) and the first level of hyponyms.

- The semantic features and ontologies integrated into the MCR, especially TO and SUMO, must always be considered.
- Glosses are usually less informative than relations and semantic features. Anyway, the annotator should pay more attention to the English glosses than to the Spanish ones since the latter might be non accurate translations of the former.

Morphology. The tagging is not performed when lemmatization mismatches appear between FreeLing and ESPWN1.6 but the case is recorded in a special separate list.

Operators. Some special notation operators had to be defined in order to allow for more flexibility in the selection of the appropriate synset. For instance we stated markers for indicating the metaphorical or metonymical use of a synset (MTF, MET), for identifying a multiword (MLTW) or for marking a partitive noun, e.g. ‘slice’ in ‘a slice of bread’ (PRT).

5.2 Structural problems

Autohyponymy. Whenever two synsets standing for the same concept are found in this relationship, the more general is chosen to tag the word –except when the context is rich enough to draw the distinction.

False hyponymy. These errors are treated case by case as separate problems of ambiguity of meaning as explained below.

5.3 Problems of ambiguity and vagueness

In cases of serious difficulty for distinguishing senses in specially complex lemmas, the judges developed a particular guide for every one. See in Figure 1 an abridged guide for ‘consejo’.

Other criteria involve specific classes of nouns, such as the following: in the case of polysemy between an objective description of an entity and the corresponding social use (e.g. between the physical and the social meaning of ‘año’-year), unless strong evidence against, the social use will be chosen.

In several cases of excessive granularity of meaning synsets have been clustered. 58 clusters affecting 129 synsets have been created.

Synsets of the lemmata ‘consejo’ :

consejo_1: it is a committee.
 consejo_2: it is a recommendation.
 consejo_3: it is about guidelines.
 consejo_4: it is also a committee.
 consejo_5: it is about a ‘tip-off’.
 consejo_6: it is a not stable committee, an improvised one.

Annotation guidelines:

- When choosing between 2, 3 and 5, sense 2 should be always used unless there are contexts where senses 2 and 5 are clearly differentiated (highly unlikely).
- 1 and 4 are virtually identical and both are hyponyms of “administrative unit”. 1 is always annotated.
- 6 is only used when it is very clear that the context deals with not stable committees. Otherwise, 1

Figure 1: Guide for disambiguating “consejo”

5.4 Problems related to incompleteness

Metaphorical and metonymic senses. When the metaphorical or the metonymic senses of a word are not declared in ESPWN1.6 we annotate the occurrence using the synset for the literal interpretation and we mark it with the MTF o MET operator.

Named Entities. Named entities (proper names, dates and amounts of money) are tagged using the MUC categories³ since they have become a standard in NLP. Moreover, we see them as well suited for establishing selective preferences – which is the final goal of SenSem.

Lack of sense. When a lemma in the corpus is not recorded in ESPWN1.6 as synset or the appropriate sense is not recorded as a variant, the fact is documented and described for a future revision of the resource.

Multiwords. When the interpretation of a multiword is compositional (e.g. ‘colegio electoral’, *electoral college*) the annotator tags analytically every part of the compound. If it is not compositional but the concept is recorded in ESPWN1.6 (e.g. ‘célula madre’, *stem cell*), it is tagged as MTW (multiword). If it is neither compositional nor is recorded in ESPWN1.6 (e.g. ‘puesta en marcha’, *the action or effect of putting on or switching on something*) the compound is tagged

³ Message Understanding Conference. http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html

using two operators: MTW (multiword) and FLT (lack of sense).

5.5 Interlinguistic problems

Interlinguistic drawbacks such as lacking Spanish equivalents or synsets split because of morphological mismatches can not be solved without changing ESPWN1.6 as it is now. Therefore the annotator is instructed to record the case for the future.

6 Results

As a result of the research here presented, 23,307 forms for 3,693 noun lemmas of the SenSem corpus have been semantically annotated with ESPWN1.6; this corresponds to the 82.6% of the total amount of verbal arguments in the corpus. In this phase of the project, in order to achieve the optimal cost-effectiveness only lemmas which occur more than 5 times have been annotated. Therefore, 17.4% of the corpus remains untagged. 91 lemmas were not tagged because they are not recorded in ESPWN1.6, being most of them culturally-specific concepts; as explained, they have been recorded for their inclusion in future versions of the Spanish WordNet.

Conclusions and future work

This paper has presented the methodology and development of a project for the semantic disambiguation and annotation of the arguments of the nominal heads of the SenSem corpus. SenSem is a balanced corpus containing 100 sentences for each of the 250 most frequent verbs in Spanish.

The result, added to previous developments in SenSem, is a richly tagged corpus both syntactically and semantically as the annotation includes verb sense, head sense of the arguments, type of phrase, argument functions, thematic roles, construction type and aspectual information. The SenSem corpus is mapped to a database bearing relevant information for each verb sense, therefore the result is a well suited resource for empirical studies focused on verbs and for the acquisition and representation of selectional preferences.

As a collateral result of the process, a critical assessment of ESPWN1.6 has been presented. Possibly, such an assessment is applicable to other WordNets as resources for annotating semantically corpus of other languages. As a result of the assessment, an annotation guide has been devel-

oped which may also be useful for similar projects.

Besides, the casuistry detected and recorded during the annotation is now being applied by our research group for developing the Spanish WordNet 3.0 (Fernandez et al., 2008; Oliver and Climent, 2011).

The SenSem corpus is freely available under a GPL license⁴.

Acknowledgements

This research has been carried out thanks to the projects FFI2008-02579-E/FILO and KNOW2 TIN2009-14 715-C04 from the Spanish Ministry of Education and Science.

References

- Eneko Agirre, Izaskun Aldezabal, Jone Etxeberria, Eli Izagirre, Karmele Mendizabal, Eli Pociello, and Mikel Quintian. 2006. Improving the Basque WordNet by corpus annotation. In *Proceedings of the Third International WordNet Conference*, pages 287-290.
- Eneko Agirre and Philip Edmonds, editors. 2007. *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*. Springer-Verlag.
- Laura Alonso, Joan Antoni Capilla, Irene Castellón, Ana Fernández, and Glòria Vázquez. 2007. The SenSem Project: Syntactic-Semantic Annotation of Sentences in Spanish. In Nikolov, K., Bontcheva, K., Angelova, G. & Mitkov, R. (Eds.), *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005*. Benjamins Publishing Co, pages 89-98.
- Javier Álvarez, Jordi Atserias, Jordi Carrera, Salvador Climent, Egoitz Laparra, Antoni Oliver, and German Rigau. 2008. Complete and Consistent Annotation of WordNet Using the Top Concept Ontology. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, pages 1529-1534.
- Ju. D. Apresjan. 1973. Regular Polysemy. *Linguistics*, 142(12):5-32.
- Jordi Atserias, Rigau G., Villarejo L. 2004. *Spanish WordNet 1.6: Porting the Spanish Wordnet across Princeton versions*. LREC'04. ISBN 2-9517408-1-6. Lisboa, 2004.
- Jordi Atserias, Lluís Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen. 2004. The MEANING Multilingual Central Repository. In *Proceedings of the Second In-*

⁴<http://grial.uab.es/SenSem/download/main>

- ternational WordNet Conference (GWC'04), pages 23-30. Brno, Czech Republic.
- Baker, C. Fillmore, and J. Lowe. 1997. The berkeley framenet project. In COLING/ACL'98, Montreal, Canada.
- Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting Parallel Texts in the Creation of Multilingual Semantically Annotated Resources: The MultiSemCor Corpus. In *Natural Language Engineering. Special Issue on Parallel Texts*, 11(3):247-261.
- Jordi Carrera, Irene Castellón, Salvador Climent, and Marta Coll-Forit. 2008. Towards Spanish verbs' selectional preferences automatic acquisition. Semantic annotation of SenSem corpus. In *Proceedings of the 6th international conference on Language Resources and Evaluation (LREC 2008)*, pages 2397-2402.
- Malcolm Davies. 2002. Un corpus anotado de 100.000.000 de palabras del español histórico y moderno. In *Procesamiento del Lenguaje Natural*, 29, pages 21-27.
- Ana Fernández-Montraveta, Glòria Vázquez, and Christiane Fellbaum. 2008. The Spanish Version of WordNet 3.0. In Storrer, A., Geyken, A., Siebert, A., & Würzner, K.M. (Eds.), *Text Resources and Lexical Knowledge*. Mouton de Gruyter, pages 175-182.
- José M. García-Miguel and Francisco J. Albertuz. 2005. Verbs, semantic classes and semantic roles in the ADESSE project. In Erk, K., Melinger, A. & Schulte im Walde, S. (Eds.), *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pages 50-55.
- Nicola Guarino. 1998. Some Ontological Principles for Designing Upper Level Lexical Resources. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 527-534.
- George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 240-243.
- Ian Niles and Adam Pease. 2003. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Model Ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering*, pages 23-26.
- Antoni Oliver and Salvador Climent. 2011. Construcción de los WordNets 3.0 para castellano y catalán mediante traducción automática de corpus anotados semánticamente. In *Proceedings of the 27th Conference of The Spanish Society For Natural Language Processing (SEPLN)*
- Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. FreeLing 2.1: Five Years of Open-source Language Processing Tools. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 931-936.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2003. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71-106.
- Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the 6th conference on International Language Resources and Evaluation (LREC 2008)*, pages 96-101.
- Piek Vossen. 1998. Introduction to EuroWordNet. Ide, N., Greenstein, D. & Vossen, P. (Eds.), *Computers and the Humanities. Special Issue on EuroWordNet*, 32(2):73-89.
- Liang-Chih Yu, Chung-Hsien Wu, and Eduard Hovy. 2007. OntoNotes: Corpus Cleanup of Mistaken Agreement Using Word Sense Disambiguation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008) Vol. I*

A Study of the Sense Annotation Process: Man v/s Machine.

Arindam Chatterjee, Salil Joshi, Pushpak

Bhattacharyya

IIT Bombay

Powai,

Mumbai, 400076.

{arindam, salilj, pb}@cse.iitb.ac.in

Diptesh Kanojia

Gautam Budh Technical University

Lucknow,

Uttar Pradesh, 226021

dipteshkanojia@gmail.com

Abstract

Does context help determine sense? This question might seem frivolous, even preposterous to anybody sensible. However, our long time research on Word Sense Disambiguation (WSD) shows that in almost all disambiguation algorithms, the sense distribution parameter $P(S/W)$, where P is the probability of the sense of a word W being S , plays the deciding role. The widely reported accuracy figure of around 60% for all-words-domain-independent WSD is contributed to mainly by $P(S/W)$, as one ablation test after another reveals.

The story with human annotation is different though. Our experience of working with human annotators who mark with WordNet sense ids, general and domain specific corpora brings to light the interesting fact that producing sense ids without looking at the context is a heavy cognitive load. Sense annotators do form hypothesis in their minds about the possible sense of a word ('most frequent sense' bias), but then look at the context for clues to accept or reject the hypothesis. Such clues are minimal, just one or two words, but are critical nonetheless. Without these clues the annotator is left in an indecisive state as to whether or not to put down the first sense coming to his mind. The task becomes all the more cognitively challenging, if the senses are fine grained and seem equally probable. These facts increase the annotation time by a factor of almost 1.5.

In the current paper we explore the dichotomy that might exist between machines and humans in the way they determine senses. We study the various parameters for WSD and also the sense marking behavior of human sense annotators. The observations, though not completely conclusive, establish the need for context for humans and that for accurate sense distribution parameters for machines.

1 Introduction

The process of sense annotation of words with senses is more accurate for humans than machines. The deciding parameter in the human sense disambiguation process is contextual evidence. Considering the principle of *weak AI*, the annotation procedure employed by the machine should make use of contextual evidence for disambiguation purposes in some form, which also conforms to the classical definition of WSD.

Our motivation is to exhibit that contextual evidence is a necessary attribute for the human tagging process. Without contextual information the human tagging process is crippled. Machines, which use the $P(S/W)$ statistic for WSD, take human context-sensitive information to learn the $P(S/W)$ measure. This is an adaptation of the contextual evidence used by human beings. Hence the principle of *weak AI* (Searle, 1980) holds for such WSD algorithms. Hence obtaining the $P(S/W)$ values perfectly is of paramount concern for machines.

A glimpse at the history of the WSD task, reveals that the initial attempt was made towards WSD in the 1980s, when machine readable knowledge resources started becoming available, especially the Princeton WordNet (Fellbaum, 1998). In this period, context-based knowledge formed the sole tool for sense disambiguation purposes. In the 1990s statistical methods gained momentum, and till date have high accuracies in the sense disambiguation process (Ide and Véronis 1998).

Today, supervised approaches to WSD deliver far better results, compared to knowledge-based or unsupervised methods (Navigli, 2009). In a supervised framework, WSD is considered as a classification task, where senses of words are the classes. If we take a closer look at the state-of-the-art supervised algorithms for WSD, it will be evident that the parameters used by

such algorithms are mostly statistical, *i.e.*, *corpus-based evidence*.

WSD researchers have tried to incorporate contextual support in the form of syntactical features, co-occurrence statistics and so on, but these algorithms do not perform significantly better over the Most Frequent Sense baseline.

A study of human cognition techniques in the annotation task unfolds that *context-based evidence* is a major parameter used by humans during annotation. In order to establish this, we made a study of the cognition techniques used by skilled lexicographers during the annotation task.

Consequently, state-of-the-art WSD algorithms use the $P(S/W)$ statistic for annotation. In this paper we attempt to answer two basic questions regarding the annotation techniques of man and machine:

For Humans: *Can humans annotate data efficiently without contextual evidence?*

For Machines: *Do machines need context information during the annotation process?*

By providing relevant answers to these questions we intend to present a comparative study of methods employed by humans and machines for sense annotation.

The remainder of this paper is organized as follows. In section 2 we present related work. Section 3 presents the different corpora and annotation scenarios used in our experiments, followed by section 4 which discusses the supervised WSD algorithm which we use as a representation of the machine annotation strategy. Section 5 describes the need for a critical analysis of the algorithm and how the analysis is done. In section 6 we provide a layout of the experimental setup, followed by the results obtained and discussions in sections 7 and 8. Section 9 concludes the paper and section 10 pertains to future-work.

2 Related Work

In this paper we compare the annotation processes of state-of-the-art algorithms, which use the $P(S/W)$ statistic as a classifier. (Lee, Ng and Chia 2004), (Khapra et. al 2010) are some examples of such state-of-the-art algorithms.

Unfortunately enough, our work seems to be a first of its kind, as to the best of our knowledge we do not know of any such work done before in the literature.

3 Corpora and Annotation Scenarios

Before elaborating on the details of the algorithm employed by us and the experiments conducted, it is essential to lay down an account of the types of corpora that have been used for our experiments and correspondingly the tagging techniques employed in each case.

It must be noted that from a linguistic point of view, the term *context* means *a set of surrounding words*. This can be a paragraph, sentence or a number of neighboring words depending on the need and focus of the experiment. In our case, we have considered the sentence surrounding the word as the context.

3.1 Context Sensitive Scenario

In this setting, a team of two skilled lexicographers was assigned the task of annotating the corpora from two *specific* domains (TOURISM and HEALTH) and a *generic* domain (NEWS), *using the context* of each word. This is a usual annotation scenario, where the lexicographer can *sense the context* and tag the word accordingly.

In order to enquire into the importance of context in the annotation processes of both human and machines, a scenario independent of contextual information was actuated.

3.2 Context Agnostic Scenario

In this setting, the same team of lexicographers was assigned the task of annotating the same corpora *without using the context*. To make the process more interesting and ensure genuineness, the corpora used in this case consisted of the list of unique words, obtained from the corpora used in the context *sensitive* scenario. The focus here was to make the lexicographers *agnostic* of the context.

3.3 Importance of Context in Annotation

After the annotation process, the lexicographers opined that this annotation task was cognitively taxing in the context agnostic scenario, which is a strong indication that context is the lone ingredient in the human annotation procedure.

In the case of machines, high accuracy WSD algorithms are mostly supervised and use the $P(S/W)$ statistic for annotation. Besides, the $P(S/W)$ statistic is obtained after training on a corpus in the context sensitive setting. Hence there is an absorption of contextual information in the generation of the $P(S/W)$ values from the context sensitive training data.

4 WSD Algorithm: Iterative WSD

In order to compare the annotations of human and machine, the machine output WSD algorithm is necessary. For our experiments we have taken the output of a supervised WSD algorithm, developed at IIT Bombay, called *Iterative WSD (IWSD)* (Khapra et al. 2010). The algorithm is greedy and uses a scoring function to disambiguate senses. The scoring function, the parameters based on which IWSD has been designed and the basic algorithm are described in the following subsections.

4.1 Parameters for IWSD

Khapra et al. (2010) proposed a supervised algorithm for domain-specific WSD and showed that it beats the most frequent corpus sense and performs on par with other state-of-the-art algorithms like Personalized PageRank (Agirre, 2009). The various parameters used by Iterative WSD can be classified as:

Wordnet-dependent parameters

- *belongingness-to-dominant-concept*
- *conceptual-distance*
- *semantic-distance*

Corpus-dependent parameters

- *sense distributions*
- *corpus co-occurrences*.

4.2 Scoring function for IWSD

The scoring function of the IWSD algorithm integrates the WordNet-dependant parameters and the corpus-based parameters to rank the candidate senses of the target word. The scoring function is illustrated below:

$$S^* = \arg \max_i \theta_i V_i + \sum_{j \in J} W_{ij} V_i V_j \quad (1)$$

Where,

$J = \text{Set of disambiguated words}$

$\theta_i = \text{BelongingnessToDominantConcept}(S_i)$

$V_i = P(S_i/\text{word})$

$W_{ij} = \text{CorpusCooccurrence}(S_i, S_j) *$

$1/WNConceptualDistance(S_i, S_j) *$

$1/WNSemanticGraphDistance(S_i, S_j)$

4.3 Algorithm

As stated earlier, IWSD is a greedy algorithm. The greedy nature of the algorithm can be ex-

plained through the steps followed by the algorithm.

Algorithm 1: *performIterativeWSD (sentence)*

1. Tag all monosemous words in the sentence.
2. Iteratively disambiguate the remaining words in the sentence in increasing order of their degree of polysemy.
3. At each stage select that sense for a word which maximizes the score given by Equation 1

Monosemous words are used as the seed input for the algorithm but are not considered while calculating the precision and recall values. It is quite possible that a sentence may not contain any monosemous words in which case the algorithm will first disambiguate the least polysemous word in the sentence. In this case, the disambiguation will be performed only using the first term in the formula which represents the corpus bias (the second term will not be active as there are no previously disambiguated words).

The least polysemous word thus disambiguated will then act as the seed input to the algorithm. IWSD is clearly greedy. It bases its decisions on already disambiguated words, and ignores completely words with higher degree of polysemy. For example, while disambiguating bisemous words, the algorithm uses only the monosemous words and ignores completely the trisemous words and higher order polysemous words appearing in the context. This is illustrated in Figure 1.

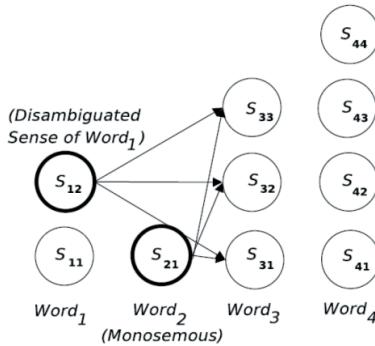


Figure 1: IWSD Operation: Only previously disambiguated words and monosemous words are used while disambiguating Word3

	Polysemous words	Monosemous words	Wordnet Polysemy	Corpus Polysemy
Noun	72225	61682	3.03	1.82
Verb	26436	4372	4.47	3.00
Adj	15462	30122	2.68	2.03
Adv	12907	10658	2.52	2.11
Overall	127030	106834	3.13	2.02

Table 1: Corpus statistics for NEWS domain

5 Critique of IWS

The accuracy of the IWS algorithm is comparable to other state-of-the-art supervised algorithms. Also IWS has both statistical as well as contextual parameters in its scoring function. To get a deeper understanding of which parameters in the IWS scoring function contribute towards its high accuracy, we performed the following tests.

5.1 Experiments on IWS

First we conducted an ablation test on the parameters of the IWS scoring function, tested on a generic corpus (NEWS), the details of which are given in table 1. The results of the ablation test are shown in table 2.

Ablation Parameter	Precision	Recall	F-Score
θ	79.61%	78.62%	79.11%
$P(S/W)$	59.59%	58.84%	59.21%
<i>Corpus-Cooccurrence</i>	79.57%	78.58%	79.07%
<i>ConceptualDistance(S_i, S_j)</i>	79.50%	78.51%	79.01%
<i>SemanticSimilarity(S_i, S_j)</i>	79.61%	78.62%	79.11%

Table 2: Results for ablation tests

We also compared the accuracy of IWS against the *Most Frequent Sense (MFS) baseline*, on the NEWS corpus. MFS tags the words based on their $P(S_i/\text{word})$ values. The results are shown in table 3. Next, in order to find the output of IWS by assigning varying weights to the statistical and contextual parts of its scoring function, we tweaked the IWS scoring formula into a linear combination of the statistical parameters and contextual parameters as explained in Equation 2, and tested for varying values of α . Table 4 shows the results of this experiment.

$$S^* = \alpha \arg \max_i \theta_i V_i + (1 - \alpha) \sum_{j \in J} W_{ij} V_i V_j \quad (2)$$

sense in almost all cases, the accuracy will be bounded as follows:

	Precision	Recall	F-Score
MFS	79.57	78.52	79.04
IWS	79.61	78.62	79.11

Table 3: MFS v/s IWS

Alpha (α)	Precision	Recall	F-score
0	59.59%	58.84%	59.21
0.00001	79.48%	78.49%	78.98
0.0001	79.50%	78.51%	79
0.001	79.50%	78.51%	79.01
0.01	79.61%	78.62%	79.01
0.1	79.61%	78.62%	79.11
0.2	79.61%	78.62%	79.11
0.25	79.61%	78.62%	79.11
0.5	79.61%	78.62%	79.11
0.75	79.61%	78.62%	79.11
1	79.59%	78.60%	79.1

Table 4: IWS results over range of alpha values

The results of tables 3 strongly indicate that IWS algorithm is marginally better than MFS. From tables 2 and 4 it is evident that $P(S/W)$ is the prime parameter for IWS.

The fact to be noted here is that, even though the $P(S/W)$ statistic apparently seems context agnostic, but it cannot be ignored that this statistic is in fact learned from corpus which was annotated in a context sensitive fashion. Hence in a way, the $P(S/W)$ parameter in IWS is an adaption of Human Context Sensitive annotations.

5.2 Accuracy estimation of IWS

Consider a sample word W which appears N times in the corpus. If W has k senses, $S_1^w, S_2^w, S_3^w, \dots, S_k^w$ which occur in the corpus with probabilities, $P_1^w, P_2^w, P_3^w, \dots, P_k^w$, respectively.

As W occurs N times in the corpus, the total no. of occurrences of S_i^w in the corpus, can be captured in the following formulation:

$$\#S_i^w = P_i^w * N$$

For an algorithm like IWS, which tags all the occurrences of a word W with the most frequent

$$\Rightarrow \min_i \{P_i^w\} \leq \text{accuracy} \leq \max_i \{P_i^w\}$$

$$\Rightarrow (1 - \min_i \{P_i^w\}) \geq (1 - \text{accuracy}) \geq (1 - \max_i \{P_i^w\})$$

$$\Rightarrow (1 - \min_i \{P_i^w\}) \geq \text{error} \geq (1 - \max_i \{P_i^w\})$$

Now, let S_d^w be the most frequent sense for W with probability P_d^w .

$$\Rightarrow P_d^w > P_i^w, \forall i \neq d$$

$$\Rightarrow \max_i \{P_i^w\} = P_d^w.$$

$$\Rightarrow \% \text{ error} \geq (1 - P_d^w) * 100$$

Since we have N occurrences of the word W in the corpus, the number of occurrences which will get tagged incorrectly will be at least,

$$N * (1 - P_d^w) \quad (3)$$

6 Experimental Setup

We report annotation experiments which were run on two specific domains (TOURISM and HEALTH) and a generic domain (NEWS). The TOURISM and HEALTH corpora consisted of around 8,000 words each and the NEWS corpus consisted of around 7,000 words.

To compare the annotated data obtained through the techniques described in Fig 1, we used *Jaccard's similarity coefficient* and *Cohen's Kappa coefficient*. We conducted the following experiments.

- We compared IWSD, Human Context Agnostic (HCA) and Human Context Sensitive (HCS) annotations taking HCS as the gold standard.
- We similarly compared the annotation genres mentioned above from the POS and ontological perspectives.

As described in section 3, we conducted experiments in both *context sensitive* and *context agnostic* scenarios to compare the annotation processes of man and machine. Sections 4 and 5 establish the algorithmic foundations of the IWSD. We can now categorize human and machine tagging into the genres illustrated in Fig 2.

Type of Experiment	Domain	POS Category				
		NOUN	ADJ	ADV	VERB	OVERALL
IWSD v/s HCA	TOURISM	0.34	0.13	0.05	0.31	0.27
	HEALTH	0.26	0.16	0.30	0.29	0.24
	NEWS	0.25	0.04	0.24	0.19	0.17

Table 5: Cohen's Kappa statistics for IWSD v/s Human Context Agnostic tagging

Human Context Agnostic (HCA)	Human Context Sensitive (HCS)
Machine Context Agnostic (MCA)	Machine Context Sensitive (MCS)

Figure 2: Human and Machine Tagging genres

7 Observations

7.1 Part-of-Speech (POS)-based and overall similarity measure

The similarity measures for Tourism, Health and News domains were calculated using Jaccard's similarity coefficient for all POS categories for every pair of annotation process as well as for IWSD. The Cohen's Kappa statistic was also calculated between Human Context Agnostic tagging and IWSD results. The results are summarized in tables 5 and 10.

7.2 Ontology-based similarity measure

Using Jaccard's similarity coefficient, the similarity measures for Tourism, Health and News domains were calculated for ontological categories for every pair of annotation process as well as for IWSD. We report the statistics for top few ontological categories that occur highest number of times. The results are summarized in tables 7 to 9.

8 Discussion

From the observations in the previous section, we see that for both the specific as well as generic domains, the similarity coefficients between pairs of annotations processes follows similar behavior. The similarity between IWSD and Human Context Sensitive tagging is highest among all three annotation comparisons. The lowest similarity occurs in case of Human Context Agnostic and Human Context Sensitive annotation pair. This behavior across all domains can be visualized as follows:

We observe similar behavior in the POS based as well Ontology based similarity measures, and the results are summarized in tables 5 to 10. The similarity measure is calculated using Human Context Sensitive data as gold standard.

We observed that the similarity between Human Context Agnostic annotations and Human Context Sensitive annotations is low across all domains, POS categories and ontological categories, which clearly indicate that the accuracy of human annotations get degraded significantly when humans try to annotate data without having knowledge of the context. Further, we also observed that extent of similarity between Human Context Agnostic annotations and Human Context Sensitive annotations was around 50%-60% across all domains, which is close to Wordnet First Sense baseline reported for these domains by Khapra et. al (2010).

Type of Experiment	TOURISM					HEALTH					NEWS				
	NOUN	ADJ	ADV	VERB	OVERALL	NOUN	ADJ	ADV	VERB	OVERALL	NOUN	ADJ	ADV	VERB	OVERALL
IWSD v/s HCA	0.72	0.56	0.61	0.74	0.68	0.69	0.56	0.79	0.69	0.67	0.66	0.40	0.75	0.53	0.62
HCS v/s IWSD	0.80	0.71	0.82	0.78	0.78	0.81	0.80	0.88	0.60	0.81	0.84	0.77	0.86	0.70	0.80
HCS v/s HCA	0.65	0.48	0.63	0.69	0.61	0.64	0.45	0.76	0.73	0.61	0.57	0.27	0.74	0.26	0.50

Table 6: Jaccard Similarity coefficient for highest occurring ontological categories across all pairs of annotation processes for HEALTH domain

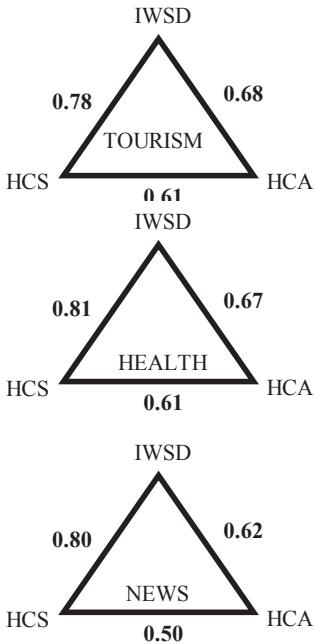


Figure 3: Visualization of comparison between tagging genres

In section 5, we have established that the prime parameter for IWSD is the $P(S/W)$ statistic. Furthermore, we have also shown that IWSD is not truly context agnostic. The similarity measure is highest for IWSD and Human Context Sensitive annotations, which is consistent with our claim.

It can also be observed that similarity between Human Context Agnostic annotations and IWSD is low compared to similarity between Human Context Agnostic annotations and Human Context Sensitive annotations, which indicates that human beings are crippled without the context based knowledge during annotation. Conversely for machines, once training is done and the $P(S/W)$ statistic is obtained, there is no further need of contextual evidence during annotation.

Ontological Category	TOURISM			
	Count	IWSD v/s HCA	HCS v/s IWSD	HCS v/s HCA
Verb of State	972	0.43	0.95	0.33
Action	863	0.25	0.83	0.21
Anatomical	798	0.35	0.89	0.34
Relational	721	0.33	0.75	0.18

Table 7: Jaccard Similarity coefficient for highest occurring ontological categories across all pairs of annotation processes for TOURISM domain

Ontological Category	NEWS			
	Count	IWSD v/s HCA	HCS v/s IWSD	HCS v/s HCA
Physical Place	2209	0.67	0.92	0.73
Person	1829	0.47	0.90	0.70
Artifact	1796	0.27	0.85	0.61
Bodily action	1582	0.20	0.83	0.55

Table 8: Jaccard Similarity coefficient for highest occurring ontological categories across all pairs of annotation processes for HEALTH domain

Ontological Category	HEALTH			
	Count	IWSD v/s HCA	HCS v/s IWSD	HCS v/s HCA
Bodily action	1198	0.06	0.95	0.89
Quantity	1188	0.01	0.90	0.16
Qualitative	1118	0.22	0.86	0.17
Numerical	1000	0.42	0.99	0.43

Table 9: Jaccard Similarity coefficient for highest occurring ontological categories across all pairs of annotation processes for NEWS domain

9 Conclusion

Based on our study on the two annotation scenarios in the two specific domains (TOURISM and HEALTH) and generic domain (NEWS), we conclude the following:

- Contextual information is paramount for humans while disambiguating sense of a word.
- The annotation process of tagging without the context is cognitively strenuous and time consuming as compared to tagging with help of the context.

- c) In the case of machines, the $P(S/W)$ measure can fetch high accuracies, provided that it has been correctly captured in the corpus by human beings, during annotation process. This in turn necessitates annotations with the help of context.
- d) WSD algorithms, if trained on corpus generated through Context Agnostic annotation process, would result in low accuracies, as the $P(S/W)$ parameter is not efficiently captured in this case.
- e) Once the training process is over and $P(S/W)$ statistic is captured, machines do not require further contextual information while annotating, unlike human annotation process. From this perspective, machines do not ape the human annotation technique but, through an adaptation of this technique provide high accuracies. Hence, machines conform to the principle of *weak AI* with respect to the annotation process.

10 Future-Work

In case of machines, we have observed that the $P(S/W)$ statistic is the machine's adaption of human context sensitive annotation process and the principle of weak AI is satisfied here. However, the accuracies for WSD algorithms are not yet at par with human annotation quality. For this, we would like to see if using better contextual parameters in the IWSB scoring function and ranking the senses using a balanced formulation between statistical and contextual parameters, can further enhance the accuracy of the machine's annotation process.

In case of humans, a deeper insight into the exact cognitive processes which are involved during the annotation process could further leverage the study between man v/s machine sense annotation processes.

References

- Eneko Agirre, and Soroa Aitor. 2009. *Personalizing pagerank for word sense disambiguation*. In EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pages 33–41, Morristown, NJ, USA. Association for Computational Linguistics.
- Christian Fellbaum, 1998. *WordNet: An Electronic Lexical Database*
- Nancy Ide and Jean Véronis. *Word Sense Disambiguation: The State of the Art Computational Linguistics*, 1998, 24(1). 2

Mitesh Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2010. *Domain-specific word sense disambiguation combining corpus based and wordnet based parameters*. In 5th International Conference on Global Wordnet (GWC2010).

Roberto Navigli. *Word sense disambiguation: A survey*. ACM Comput. Surv. 41, 2 (2009).

Lee, K. Yoong, Hwee T. Ng, and Tee K. Chia. 2004. *Supervised word sense disambiguation with support vector machines and multiple knowledge sources*. In Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pages 137–140.

Searle John. 1980. *Minds, brains, and programs*, Journal of Behavioral and Brain Sciences, Vol. 3, pages 417-424.

A Computer Aided Approach for Enriching WordNet with Semantic Definition

Lin Dai

Weitao Zhou

Heyan Huang

Beijing Institute of Technology

5 South Zhongguancun Street, Haidian District, Beijing 100081, China.

{dailiu,1212, hhy63}@bit.edu.cn

Abstract

WordNet is studied and applied extensively nowadays as an underlying knowledge base. The semantic information in WordNet is an important evidence for many applications. Because the synsets are defined in natural language, which is difficult for computer to understand, it is meaningful to extend it with computer-readable semantic definitions. Motivated by this objective, this paper proposed a computer aided method to give synsets semantic definitions. These semantic definitions are picked up from HowNet, an English-Chinese bilingual Ontology, in which the definitions can be easily utilized by computers and thus will make WordNet more attractive. Experimental results show that the proposed method is capable of this enriching task.

1 Introduction

WordNet¹ is a lexical database for the English language(Miller et al., 1990). It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. So far, WordNet has been used for many of different purposes in information systems, such as word sense disambiguation (Pedersen et al., 2004), information retrieval (Tokunaga et al., 1998) (Gonzalo et al., 1998), automatic text classification (Scott and Matwin, 2009) (Gonzalo et al., 2009), automatic text summarization (Chaves, 2001), and so on. Furthermore, WordNet can also be widely interpreted and used as a lexical ontology in the computer

science sense. For example, Mann (2002) explored the idea of a fine-grained proper noun ontology based on WordNet in question answering. Buscaldi (2005) described a method of using the WordNet ontology in the GeoCLEF geographical information retrieval task.

As well known, WordNet organizes concepts into hierarchies and gives each synset a natural language definition and usage samples. By exploring these information, computer based systems can perform semantic computing. But because the definitions and samples are given in natural language, it is easy for people to understand but difficult for computers to do so. Thus it is very meaningful to give each synset a computer readable definition. This is the objective of this paper.

Besides Wordnet, stimulated by monolingual and multilingual NLP applications, some WordNet-style or non-WordNet-style ontologies have been constructed. HowNet (Dong, 2003)², which reveals the relationship among concepts, and the relationship among attributes of concepts, is such a non-WordNet-style knowledge base. HowNet takes a constructive approach to building a lexical hierarchy. HowNet defines each concept with a definition expression, which can be easily utilized by computers. At the most atomic level of definitions, there is a set of about 2000 basic semantic definitions (or meta-concepts), such as “*human*”, or “*bird*” (Carpuat, 2002). In HowNet, relations between concepts are represented by their definitions. HowNet’s special definition system and design philosophy facilitate many NLP applications such as automatic text summarization, word sense disambiguation and so on (Dai et al., 2008).

Because both HowNet and WordNet are describing concepts of this world, there are many

¹WordNet 3.0 is used in this paper.

²HowNet version 2010 is used in our experiments.

identical concepts in them. It is possible to transfer the definitions of the these concepts from HowNet to WordNet. Nevertheless, manual alignment between WordNet's synsets and HowNet's concept definitions would be extremely expensive. In this paper, we propose a computer aided method for this task. The key part of the method is to rank the candidate definitions according to relevance by using Latent Semantic Analysis (LSA) technique(Landauer et al., 1998).

The remainder of this paper is organized as follows. Section 2 briefly describes the definition system of HowNet. The proposed method is described in Section 3. Section 4 describes the relevance ranking algorithm of candidate definitions. And Section 5 presents the experimental results before concluding this paper in Section 6.

2 Definition system of HowNet

In this section we briefly introduce the definition system of HowNet, which is firmly related to our task.

As pointed out by its creator (Dong, 2003), "HowNet is an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts". The philosophy behind HowNet grounds its understanding and interpretation of the objective world. The crux is that all matters (physical and metaphysical) differentiate themselves from others by their attribute(s).

As a knowledge base for natural language processing, HowNet could be viewed as firstly a general ontology, secondly a thesaurus, and lastly a dictionary. It provides plenty of taxonomically semantic knowledge as well as real-world knowledge. As a knowledge system that describes relations between concepts, HowNet is not a traditional thesaurus. HowNet attempts to construct a graph structure of its knowledge base from the inter-concept relations and inter-attribute relations. This is the fundamental distinction between HowNet and other tree-structured lexical databases. The philosophy of HowNet and its nature entail its unique structure.

The meta-concept in HowNet is called *sememe*, which is the smallest semantic unit that can't be reduced further and used to define other concepts.

³The current version of HowNet is bilingual, i.e., English and Chinese. Sememes are labeled with English words and Chinese words at the same time. The English sememe set is essentially independent and redundant to the Chinese sememe set. Actually, both English words and Chinese words can be defined by solo-lingual sememe set. In other words, both English version and Chinese version of sememe set is competent to define words of different language. Thus, we only give the English version of the sememes.

Through combining sememes selected from a finite sememe set, HowNet can describe infinite concepts. There are 2219 basic sememes in the current version of HowNet which are organized into five taxonomies, namely *Entity*, *Event*, *Attribute*, *Attribute Value*, and *Secondary Feature*. Besides the basic sememes, HowNet uses subsidiary sememes to describe the world.

As the sememes are concerned, in addition to being fundamental description units, there are complicated relations among them. The relations include hypernym-hyponym, antonym, converse, whole-part, material-product, attribute-owner, event-roller, etc. In HowNet, these relations are given in two ways. One way is the tree structure which represents hypernym-hyponym in taxonomies. Figure 1³ is a part of the tree structure of the *Entity* taxonomy. The other way of giving relations is the definition frame of sememes. Although sememes are the meta-concepts used to define other concepts, they may have their own definition frame, which is again made up by sememes. For example, the frame of sememe *bird* is: {animal:materialOf={edible}, {eat:patient={~}}, {fly:agent={~}}}, and the frame of sememe *food* is: {edible:{cook: PatientProduct ={~}}}. Because both of these two frames have a sememe *edible*, we can tell that there is a strong relation between sememe *bird* and *food*. From the relations among sememes we can see that the actual structure of HowNet is indeed a net rather than merely a tree. By thoroughly exploring these relations, we may get evidences for many NLP applications.

From the relations among sememes we can see that the actual structure of HowNet is indeed a net rather than merely a tree. By thoroughly exploring these relations, we may get evidences for many NLP applications.

Although there exist tree hierarchies in HowNet, it is totally different to WordNet. In WordNet, concepts are represented by synsets and they are the minimum unit for word sense description and each concept is a node in a hierarchy. But in HowNet, sememes are nodes in hierarchies and each concept is defined by sememes.

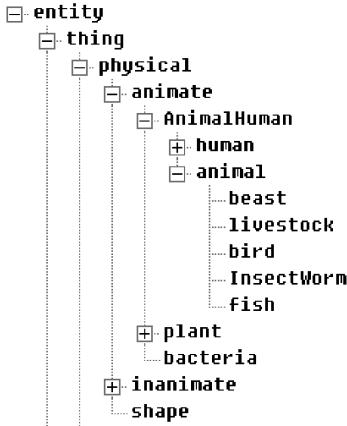


Figure 1: A part of the tree structure of the Entity taxonomy in HowNet

HowNet uses *Knowledge Database Markup Language* (Dong, 2003) (KMDL, in short) to construct the semantic expression of concepts and sememes. KMDL consists of a set of grammar rules and keywords. The detailed introduction of KMDL is beyond this work. Hence we give several definition examples in Table 1.

Word	Definition
cock	{bird:modifier={domesticated}}{male}}
fruit	{fruit}
asylum	{InstitutePlace:domain={medical}, {doctor:content={disease:CoEvent ={mad}}}.location={~}}
shore	{land:{BeNear:existent={~},partner={waters}}}
cushion	{furniture:{sit:location={~}}}
grin	{CausePartMove:PatientPart={part: PartPosition={mouth},whole={AnimalHuman}}}
jewel	{material:MaterialOf={treasure}}
middy	{time:TimeSect={afternoon}}
stove	{tool:{WarmUp:instrument={~}}, {burn:location={~}}}

Table 1: The examples of definition expression of concepts in HowNet

The current version of HowNet includes mainly frequently used English words and Chinese words, as shown in the Table 2. Its large coverage makes it competent to act as a knowledge base in many NLP tasks.

Item	Number
English word and expression	96370
English meaning	121042
Concept definition	29868
Total record	191924

Table 2: The coverage of the current version of HowNet

A complete sample of semantic definition in HowNet is shown in Figure 2.

$NO. = 023648$
 $W_C = 打(Pinyin: da)$
 $G_C = CLAS [da3]$
 $E_C = \text{十二个一打}(Pinyin: shi er ge yi da. a dozen is twelve), 一打铅笔(Pinyin: yi da qian bi. a dozen of pencils), 两打鸡蛋}(Pinyin: liang da ji dan. two dozens of eggs), 论打出售(Pinyin: lun da chu shou, sale in dozen), 买两打}(Pinyin: mai liang da. buy two dozen of)$
 $W_E = dozen$
 $G_E = N$
 $E_E =$
 $DEF = \{NounUnit\}名量(Pinyin: ming liang. quantifier); host=\{inanimate | 无生物}(Pinyin: wu sheng wu. inanimate)\}\}$

Figure 2: An example of a concept in HowNet.

In Figure 2, “ $NO.$ ” is the index of the entry. “ W_C ” is the Chinese word. “ G_C ” is the POS and the Pinyin of this Chinese word. “ E_C ” gives some examples of this Chinese word. “ W_E ” is the corresponding English word. “ G_E ” is the POS of the corresponding English word. “ E_E ” gives some example of this English word, if any. “ DEF ” is the definition of this Chinese word.

3 Basic Idea

As discussed above, WordNet gives each concept, i.e., synset, a nature language definition and usage samples. On the other hand, HowNet gives each concept a semantic definition based on meta-concepts. If synsets in WornNet have semantic definition as HowNet does, at least two benefits will be entailed: 1) both human rates and computers can easier understand synsets. 2) more useful semantic information will be available to WordNet based systems.

But it will take much labor to annotate semantic definition to synsets. So we are seeking a computer aided method to reduce the labor. This method can give a definition to given synset if possible, or it can give the candidate definitions for synsets from which people can select or generate proper definitions.

The feasibility of the method lies in the nature of WornNet and HowNet: Both of them are dedicated to define concepts of this world. For a given synset, if there is one and only one described concepts in HowNet, we can directly transfer the definition to the synset. Otherwise, the method tries to find a sorted list of candidate definitions from HowNet according to relevance.

Given a synset $synset_n$, the method consists of two steps, as follows.

Step 1. Eliciting possible definitions

The starting point is to find out all the possible matching semantic definitions for $synset_n$. For each word in the synset, we look up HowNet to find the definitions that have the same POS as this synses. All these definitions make up the synset's possible definition set. Figure 3 elaborates this step.

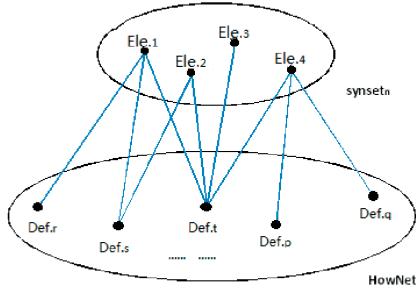


Figure 3: Example of eliciting semantic definitions to synset

In this example, each of the four words of $synset_n$ has one or more possible related concepts in HowNet. The possible definitions of words are:

```

set1: {Defr, Defs, Deft}
set2: {Defs, Deft}
set3: {Deft}
set4: {Defp, Defq}

```

The union of the possible definition sets makes up the possible definition set of the synset $\{Def_r, Def_s, Def_t, Def_p, Def_q\}$.

Step 2. Handling definitions

After getting the possible definition sets of elements of a synset, we compute the intersection of them. There should be three cases:

- **Case 1:** The intersection consists of only one definition.
- **Case 2:** The intersection is empty.
- **Case 3:** The intersection consists of more than one definition.

If the result is Case 1, the only definition of the intersection is directly deemed as the semantic definition of the synset. For example, in Figure 3, Def_t is the only resulting element in the intersection, thus it becomes the semantic definition of $synset_n$.

If the result is Case 2 or Case 3, we need to rank the possible definitions. Definitions with high rank will become the candidate definition of the synset. By this way, people can select or modify the candidate definitions to get the final definition.

4 Ranking Candidate Definitions

In order to rank candidate definitions, we compute the relevance of each candidate definition to the given synset. The available information for synsets includes nature language descriptions and examples(for example, see Figure 4). Information from HowNet includes semantic definitions and examples, i.e., $\{E_C\}$ (in Figure 2). Note that we should translate examples $\{E_C\}$ from Chinese to English for further computation.

test, trial, run – (the act of testing something; “in the experimental trials the amount of carbon was measured separately”; “he called each flip of the coin a new trial”)

Figure 4: A sample of WordNet's synset. “*test, trial, run*” are the *elements* making up this *synset*; “*the act of testing something*” is a description of this *synset*. And the last two sentences are usage examples about this *synset*.

Both the available information of HowNet's concepts and WordNet's synset makes up a short document. There is a problem that the feature space is very sparse. Under such a circumstance, standard document similarity measures will fail because they rely heavily on term counts in documents. If the short documents do not have any

terms in common, then they receive a very low similarity score, regardless of how topically related they actually are. To overcome the difficulty caused by the sparsity, Latent Semantic Analysis (LSA) is introduced to represent the short documents in this paper.

4.1 Computing Relevance with LSA

Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of document (Landauer et al., 1998). LSA is closely related to neural net models, but is based on singular value decomposition, a mathematical matrix decomposition technique closely akin to factor analysis. As a practical method for the characterization of word meaning, LSA produces measures of term-term, term-document and document-document relations that are well correlated with several human cognitive phenomena involving association or semantic similarity.

To successfully apply LSA to short document relevance measure, we need to follow the steps below:

Step 1: Feature extraction

First, a feature matrix (W) must be constructed from training data. The matrix W consists of measures of co-occurrence between term and document. This measure needs suitable functions of the number of times each term appears in each document. In this paper the (i, j) cell of W is obtained as:

$$W_{ij} = (1 + \log_2(t_{ij})) \log_2\left(\frac{C}{d_i}\right) \quad (1)$$

where t_{ij} is the number of times term(i) occurs in document(j), d_i is the document frequency(DF) of term(i), and C is the total number of documents in training data.

Step 2: Singular value decomposition(SVD)

Secondly, perform the SVD on W as follows:

$$W = T_0 \cdot S_0 \cdot D'_0 \quad (2)$$

such that T_0 and D_0 have orthonormal columns and S_0 is diagonal. T_0 and D_0 are the matrices of left and right singular vectors and S_0 is the diagonal matrix of singular values.

Once we get the Matrices T , S , D by SVD, a latent semantic space (LSS, in short) is successfully decided. Then we further get the approximate LSS with reduced model:

$$W \approx \widetilde{W} = T \cdot S \cdot D' \quad (3)$$

which is the rank- k model with the best possible least-squares-fit to W . The dimension k is bounded from above by the (unknown) rank of the matrix W , and from below by the amount of distortion tolerable in the decomposition.

Step 3: Document-Document comparisons

Documents similarity can be computed on the latent semantic space. Given a document, the following equation is used to project it to LSS :

$$E = F \cdot T \cdot S^{-1} \quad (4)$$

where F is the feature vector of a document, T is the left matrix of SVD, S is the singular values matrix of SVD, and E is the feature vector on LSS .

At last, the relevance between two feature vector of latent semantic space is computed by cosine similarity:

$$Rel(E_a, E_b) = \frac{E_a \cdot E_b'}{\|E_a\| \cdot \|E_b\|} \quad (5)$$

4.2 Ranking Definitions

To use LSA to rank candidate definitions, we needs to transform the information of HowNet's concepts and WordNet's synset into short documents. Because their structures are different, we handle them differently.

- HowNet's definitions are reorganized into short documents according to the following format:

$$\begin{array}{c|c|c} W_E & \text{sememes in } DEF & E_C \\ H & M & L \end{array}$$

Take Figure 2 for example, the resulting short document is:

$$\begin{array}{c|c|c} dozen & \text{NounUnit inanimate} & | \\ twelve & dozen, a dozen pencils, two & \\ dozen & eggs, On the sale of playing, & \\ & buy two dozen & \end{array}$$

- Meanwhile, WordNet's synsets are reconstructed into short documents according to the following format:

$$\begin{array}{c|c|c} elements & \text{description} & \text{examples} \\ H & M & L \end{array}$$

Take Figure 4 for example, the short document is:

test trial run | the act of testing something | in the experimental trials the amount of carbon was measured separately he called each flip of the coin a new trial

It is intuitive that the part H and M are more informative to relevance estimation. To embody the difference, before using equation 1 to calculate weight, features in H and M are enhanced experientially by multiply their counts, with multiplier $h=100$ and $m=80$ experientially, respectively.

So far, we can rank the candidate definitions of given synset according to their relevance to a given synset.

	# Synsets	# Concepts	# Case 1	# Case 2	# Case 3
NOUN	82115	63262	12564	35454	15244
VERB	13767	27427	2158	5479	6130
ADJ	18156	12408	3561	4841	4006
ADV	3621	2860	656	1432	772

Table 3: Statistics of three cases of intersection.

We firstly check the synsets whose definition could be decided, i.e., the intersection of definition sets of words in it contains only one element (Case 1). The accuracies and recalls are presented in table 4.

	Accuracy	Recall
NOUN	93.5	15.3
VERB	92.6	15.7
ADJ	85.1	19.6
ADV	90.0	18.1

Table 4: Accuracies and Recalls of Case 1

From Table 4 we can see that see that 1) The number of concepts in HowNet and WordNet are different with HowNet has slightly more entries in general, 2) more than 15% synsets can be successfully given a semantic definition automatically with good quality.

For the other two cases, i.e., the intersection of definition sets of words in a synset contains none or more than one element (Case 2 and Case 3), we rank the all the definitions related to the words.

5 Experiments and Results

Wikipedia database⁴ is used as LSA’s training data set. In our experiments we use the English Wikipedia database dump from 20 June 2011, which contains 3,720,790 articles covers broadly topics. The large coverage of this data set will give sound latent relevance between words. The google translation⁵ is used to translate Chinese examples of HowNet’s concepts into English.

To estimate the accuracy of the resulting definitions, in each experiment, we randomly select 300 synset-definition pairs and manually decide if it is correct or not. Then the estimated accuracy is recorded. And the recall is also recorded in each experiment.

The statistic features of the two knowledge base are shown in table 3, classified by POS.

	Accuracy	Recall
NOUN	80.5	89.3
VERB	66.6	79.7
ADJ	34.1	87.6
ADV	23.0	45.1

Table 5: Statistics on Ranking Algorithm

It can be seen from Table 5 that, in general, NOUNs and VERBs have higher accurate and recall than ADJs and ADVs. It is partially caused by the training data of LSA, because Wikipedia corpus mainly concentrates on entitative objects.

Table 6 gives two examples of ranked definitions of two synsets. It can be seen that the synset “animal, animatebeing, beast...” is firmly related to top ranked semantic def-

⁴Wikipedia can be downloaded at <http://download.wikimedia.org/enwiki/>

⁵<http://translate.google.com>

initions “livestock”, “beast” and “AnimalHuman”. Meanwhile, the synset “course, course of action” is close to semantic definitions “process”, “route:fly:agent=aircraft ” and “Se-

quence:host=event”. From these two examples we can see that the granularity of HowNet is finer than WordNet. As a conclusion, the ranked list of definitions is informative to define a synset.

Syn.		Se. Def.	
	W.E	DEF	Sim.
animal, animatebeing, beast, brute, creature, fauna	beast	{livestock}	0.95
	animal	{beast}	0.89
	fauna	{AnimalHuman}	0.89
	brute	{human:modifier={wicked}}	0.55
	creature	{human:modifier={special}}	0.20
course, course of action	course	{process}	0.94
	course	{route:{fly:agent={aircraft}}}	0.79
	course	{Sequence:host={event}}	0.77
	course	{direction:{SelfMoveInManner}}	0.55
	course	{facilities:{exercise:location={}}}	0.34
	course	{food:{cook:PatientProduct={}}}	0.11

Table 6: Two examples of ranked semantic definitions of synsets

6 Conclusion and Future Work

This paper tries to give each synset of WordNet a computer-readable semantic definition in semi-automatic manner. The semantic definitions are transferred from HowNet, a bilingual knowledge base. The proposed method can give a synset a solo definition or a ranked list of candidate definitions. The candidate definitions are ranked via LSA algorithm in latent feature space, with Wikipedia as background corpus. Experimental results show that our method is promising.

We are planning to introduce more comprehensive corpus to improve the accurate and recall of the ranking algorithm in future. We are also planning to make a English-Chinese bilingual WordNet based on this work.

Acknowledgments

This work is supported by National Natural Science Foundation of China (60803050, 61132009) and BIT Team of Innovation.

References

D. Buscaldi, P. Rosso and E.S. Arnal. 2005. Using the WordNet Ontology in the GeoCLEF Geographical

Information Retrieval Task. *CLEF’05*, Pages:939–946

M. Carpuat. 2002. Creating a Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet. In *Proceedings of the 1st Global WordNet Conference, Mysore*, Pages:284–292.

R. P. Chaves 2001. WordNet and Automated Text Summarization. In *Proceedings of the 6 th Natural Language Processing Pacific Rim Symposium, NLP-PRS*.

L. Dai, Bin Liu and etc. 2008. Measuring Semantic Similarity between Words Using HowNet. *ICCSIT ’08: Proc. of the International Conference on Computer Science and Information Technology*, pp.601–605. IEEE.

Z.D. Dong, and Q. Dong. 2003. HowNet–A Hybrid Language and Knowledge Resource. *Proceeding of the International Conference on Natural Language Process and Knowledge Engineering*, Pages:820–824.

J. Gonzalo, F. Verdejo and etc. 1998. Indexing with WordNet synsets can improve Text Retrieval. *CoRR*.

G.S. Mann. 2002. Fine-Grained Proper Noun Ontologies for Question Answering. *SemaNet’02: Building and Using Semantic Networks*, volume 1. Association for Computational Linguistics Stroudsburg, PA, USA.

- T.K. Landauer, and P.W. Foltz, and D. Laham. 1998. Introduction to Latent Semantic Analysis, *Discourse Processes*, NO. 25. Pages:259–284.
- J. Li, Y. Zhao, and B. Liu 2009. Fully Automatic Text Categorization by Exploiting WordNet. *Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology*. Springer-Verlag, Berlin, Heidelberg, pp.1-12.
- G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.
- T. Pedersen, S. Patwardhan, and J. Michelizzi. 2004. WordNet::Similarity-measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*. Association for Computational Linguistics, Stroudsburg, PA, USA, 38-41.
- S. Scott and S. Matwin 1998. Text Classification Using WordNet Hypernyms. *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*. Association for Computational Linguistics, pp.45–52.
- R. M. Tokunaga, M. Rila, T.Takenobu and Tanaka Hozumi. 1998. The Use of WordNet in Information Retrieval.

Combining Wordnet and Crosslingual multi-terminology health portal to access health information

**Stefan J Darmoni, Julien Grosjean,
Tayeb Merabti, Nicolas Griffon,**

Badisse Dahamna

CISMef & TIBS LITIS, Rouen University Hospital, Normandy, France.

{Stefan.Darmoni;
Julien.Grosjean;
Tayeb.Merabti; Nico-
las.Griffon;
Badisse.Dahamna}@chu-
rouen.fr

Dominique Dutoit

SenseGates, Vigneux, France

d.dutoit@sensagent.com

Abstract

Background: Wordnet® is a large multilingual lexical database in various languages. The European Health Terminology/Ontology Portal (EHTOP) provides access to terminologies and ontologies, allowing dynamic browsing and navigation. EHTOP can be used either by individuals or by computer systems via Web services.

Methods: To integrate terminologies and ontology into EHTOP, three steps are necessary: (1) designing a meta-model into which each terminology and ontology can be integrated, (2) developing a process to include terminologies into EHTOP, (3) building and integrating existing and new inter & intra-terminology semantic harmonization into EHTOP.

Results: EHTOP is freely available for the MeSH in French (URL: <http://pts.chu-rouen.fr>). The access to other terminologies and to other languages (mainly English and French, but also German, Italian and Greek.) is restricted and available only for the scientific community. A total of 32 terminologies are included in EHTOP, with 980,000 concepts, 2,300,000 synonyms, 222,800 definitions and 4,000,000 relations. More recently, the cross-lingual version of EHTOP has been made available (URL: http://cispro.chu-rouen.fr/ehtop_site/) and freely provides access to ICD10 in five languages.

Combining EHTOP to Wordnet has improved both tools: allowing not only for EHTOP to obtain definitions and a rich source of information but also for Wordnet to have access to more health terms, including definitions and therefore expanding its database.

Conclusion: Combined with a cross-lingual dictionary, EHTOP is a useful tool for a wide range of applications and users, whatever in education, resources indexing, information retrieval or performing audits in terminology management.

MeSH keywords: Abstracting and indexing; Cataloguing; Controlled Vocabulary; Internet; Database; Dictionaries; Europe; Information Storage and Retrieval; Internet; Subject Headings; Terminology as subject.

1 Introduction

The Internet is currently the major source of scientific and health information as well as knowledge. If health information and knowledge for health professionals is generally in English, health information for lay people is also available in each language. Moreover, as people around the world are travelling more and more, health information should transcend borders. It should become multilingual and based on different health terminologies and ontology sources. Some institutions are already proving health information in several languages: e.g. MEDLINEplus¹ is providing health information

¹ <http://www.nlm.nih.gov/medlineplus/> (August 2011)

for lay people in English and Spanish. Furthermore, Europe Medicines Agency² is providing drug information for health professionals and lay people in each European language.

Health is with law the main fields, where several terminologies and ontology (T/O) coexist. For the English languages, over 150 terminologies and classifications are included in the Unified Medical Language System (UMLS) meta-thesaurus developed by the US National Library of Medicine since 1986.

There is an increasing amount of interest today not only in developing and maintaining health-care T/O but also in making it interoperable within information technology (IT) systems delivering services to applications. Terminology server has been defined as a tool to manage and to give access to various terminologies (Burgun *et al.* 1997). Several terminology servers have already been developed, mostly in English. One was recently developed for French terminologies by three partners (Darmoni *et al.* 2009): the private company Mondeca and two academic medical informatics labs: the LERTIM from Marseilles and the CISMeF team from Rouen, France.

Wordnet® (n.d.) is a large lexical database for English which is also available in various languages, including French for this work: the SenseGates Web site (n.d.) is providing this translation. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept that should be linked to EHTOP. The main relationships between words in WordNet is synonymy (Wordnet, n.d.; Fenbaum, 2005), where in EHTOP the main relationships is the hierarchy among descriptors, although synonymy is also an important relationships (e.g. CISMeF has manually added 16,000 French synonyms to the MeSH thesaurus in the last 16 years).

Co operation between the CISMeF team and SenseGates (previously Memodata) began in 2004 with the Vodel project with the same objective “ontological valorization of electronic dictionaries” following in 2007 by the InterSTIS project³ (Darmoni *et al.* 2009) (Semantic Interoperability between T/O), both funded by the French National Research Agency. Bilingual dictionaries based on Wordnet (Alexandria tool) have been integrated in various CISMeF tools, in particular the CISMeF quality-controlled health

gateway (Catalog and Index of Health Resources in French-Fr) [URL: <http://www.cismef.org>] (Darmoni *et al.*, 2000) and French MeSH Browser (Thirion *et al.* 2007).

The aim of this work was to create a Health Multi-Terminology Cross-Lingual Portal mainly based on European languages (EHTOP) as well as combining EHTOP to WordNet®. The primary goal of EHTOP was to search concepts among all the health terminologies available in French (or in English and translated to French) included in this portal and to browse it dynamically. The ultimate goal was to use this search: (a) to index resources manually or automatically in the CISMeF quality-controlled health gateway (Catalog and Index of Health Resources in French-Fr) [URL: <http://www.cismef.org>] (Darmoni *et al.*, 2000); (b) to allow multi-terminology automatic indexing and information retrieval; (c) to evaluate the integrity of terminological data (audit); (d) to provide a new useful tool to train health students.

2 Material and methods

2.1 List of T/O included in EHTOP

Thirty two T/O were included in the CISMeF Information System (n=32), and therefore in EHTOP (see Figure 1). Some T/O are included in the UMLS meta-thesaurus (n=11) but most are not (n=21).

Among these 32 T/O:

- MeSH thesaurus (National Library of Medicine, 2010), including the MeSH Supplementary Concepts (MeSH SC), including the translation of 15,300 MeSH SC in French and the inclusion of over 16,000 synonyms to MeSH terms; the MeSH thesaurus is the pivotal thesaurus used to index health resources in the CISMeF catalogue. During the first 10 years of existence, CISMeF used only one thesaurus to index Internet resources as only MeSH is used in the MEDLINE bibliographic database. Because Internet resources are more diverse than scientific articles, the CISMeF team underwent a major strategic shift in 2005: switching from a mono-terminological world to a multi-terminology universe for the overall CISMeF information system, which includes multi-terminology automatic indexing, multi-terminology information retrieval and integration of several terminologies into

² <http://www.ema.europa.eu/> (August 2011)

³ <http://www.interstis.org/about-interstis/> (August 2011)

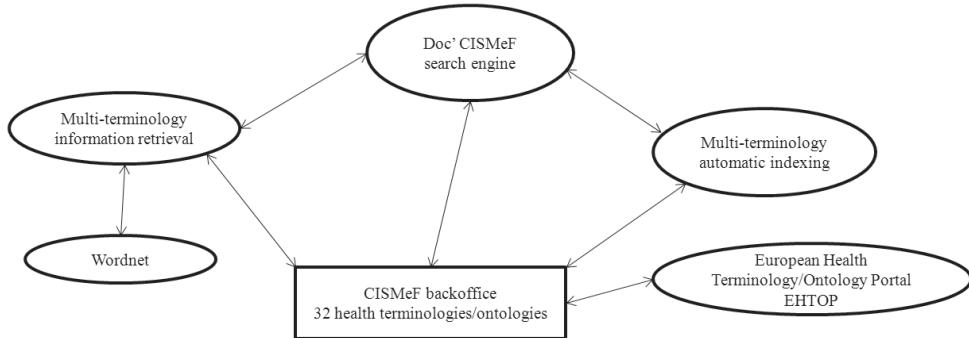


Figure 1: Interrelationship between CISMeF Information System, EHTOP and Wordnet

the CISMeF information system, then allowing the creation of the EHTOP portal.

- CISMeF thesaurus, is an extension to the MeSH thesaurus, including 130 metaterms (super-concepts to unify MeSH terms of the same medical discipline), 300 resource types (adaptation to the Internet of the publication types), over 200 predefined queries and the translation of 12,000 MeSH Scope Notes (8,000 manually and the rest semi-automatically); this semi-automatic translation was performed by Sensegates tools based on Wordnet.
- SNOMED CT & SNOMED International (French version) to describe electronic health records (Lussier, 1998),
- And all T/O developed by the World Health Organization (WHO), in particular ICD10 (International Classification of Diseases, 10th revision) (WHO, n.d.).

Some other terminologies and ontology will be integrated in the coming months, in particular US National Cancer Institute Terminology. Most of these T/O (those freely available), in particular all the MeSH add-ons were then included in Sensegates tools to improve its dictionaries with health and scientific terms and its definitions.

2.2 Integration of the terminologies

To integrate the terminologies in the CISMeF Information System (Oracle 11.1g database), three steps are necessary:

- to design a terminology generic model into which each terminology model can be integrated,
- to design a process capable of integrating terminologies into the EHTOP,
- to build and integrate intra and interterminology semantic harmonization into EHTOP. Two inter-terminology mappings were performed automatically: one based on

UMLS concepts and one based on NLP tools developed by the CISMeF team.

A generic model was designed for the database in order to fit all the terminologies into one global structure (see Figure 2): this database is the CISMeF Information System. Then, a model of each terminology was designed as a specialization of the meta-model. Many conceptual and technical issues were encountered, especially in the model creation for several terminologies (MedDRA model, FMA ontology to terminology). It was necessary to understand the whole structure and the functional purpose of each terminology to propose a good representation for users. Another problem with terminologies is the space complexity when data are very large (e.g. SNOMED International with more than 80,000 terms and 62,000 relations). We had to adapt our tools to allow integration in short period of time while keeping a control on data.

2.3 The CISMeF Information System

This system was established around the "Descriptor" which is the central concept of the terminologies (aka "keyword"). Each descriptor is labelled and may be defined, linked to other descriptors (such as Related-Term relation) and involved in a son-father type of hierarchy (BT-NT for Broader Term – Narrower Term), which is the main relation within EHTOP. A descriptor may also contain specific attributes and synonyms (which is the main relation of Wordnet), abbreviations etc.

It was also necessary to work on the terminologies modeling (OWL format) in order to fit it into the global database structure and to standardize the data in a well known and shared format. That is why the RDF (Resource Description Framework) syntax was chosen with the OWL (Ontology Web Language).

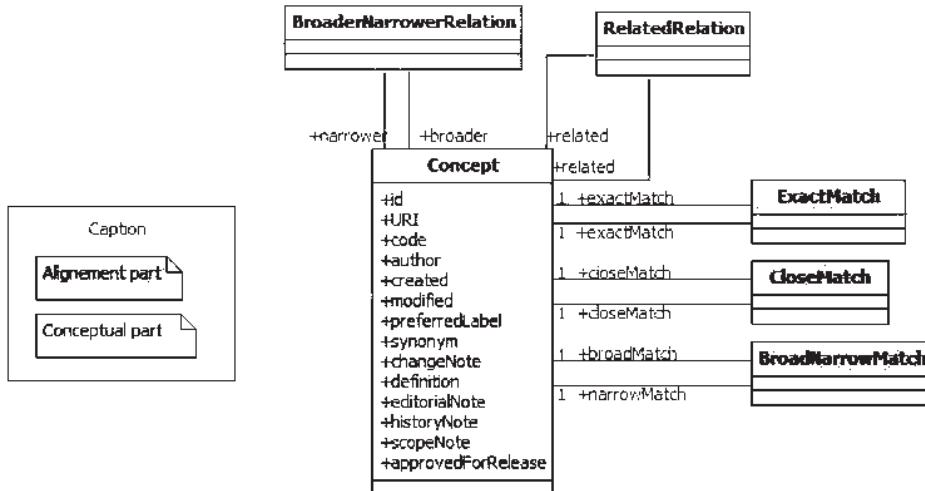


Figure 2: The CISMeF BackOffice database conceptual structure

2.4 Creation of the EHTOP

The EHTOP was designed as a graphic interface of a Web Service, entirely dedicated to information retrieval and associations between terms of several terminologies. Thus, the main objective was to dissociate the substance from the form, in particular the interface.

3 Results

This project is human-oriented work to deal with terminologies and ontologies. In the EHTOP modelization, data is more important for humans as opposed to structure for computer. Therefore, we decided to oversimplify ontologies and when necessary integrated terminologies into EHTOP. Finally, on July 2011, the time spent to build and maintain this health multi-terminology multilingual portal was estimated as to be approximately nine man-years. Currently, the CISMeF team is using one junior engineer (JG) to integrate new terminologies (e.g. SNOMED CT) and one post-doc (TM) to perform semantic harmonization on each terminology to another (almost 500 alignments were already performed -32*31/2-) using CISMeF NLP tools.

Currently, two versions of EHTOP exist: (a) the first version is mainly bilingual (French and English) specifically devoted to French users. This version is available at the following URL: <http://pts.chu-rouen.fr/>. Only MeSH and CISMeF terminologies are freely available. We also provide a restricted access only for the scientific

community (click on “Subscribe”); (b) the second version is multilingual. Because the right to access terminologies varies among countries (e.g. the MeSH is freely available in English and French, but it is not the case in other European languages), the EHTOP multilingual version provides a free access to ICD10 in five European languages (English, French, German, Dutch, Greek).

A total of 32 terminologies are included in EHTOP, with 980,000 concepts, 2,300,000 synonyms, 222,800 definitions and 4,000,000 relations. Twenty-one of these terminologies are not yet included in the UMLS including some from the World Health Organization. Due to various optimizations, the average response time for one concurrent user takes less than 500 milliseconds. The access to Wordnet content is available on each page of CISMeF (see Figure 3). The same figure displays that Wordnet content was enriched by CISMeF definitions and hierarchies. Some hyperlinks allow navigation through both interfaces.

Via its human interface, since January 2010, EHTOP is used daily by the CISMeF team to index resources in multi-terminology mode. It is also used by various CISMeF academic partners in different French and European projects, which was necessary to develop the EHTOP. EHTOP was qualitatively evaluated in 2010 by 25 medical students from the Rouen Medical School and gave 58% satisfaction for its user interface and 76% for its functionalities and content.

CISM F

Catalogue et Index des Sites Médicaux de langue Française

Information : Un double clic sur un mot permet d'en afficher la définition.

Terme en Anglais : liver abscess; amebic

Définition [Définition MeSH] : Abscès du foie provoqué par ENTAMOEBE HISTOLYTICA

Définition [Définition VIDAL] : Forme grave de l'infection par l'ameba, provoquant une sorte d'abcès du foie.

Synonyme CISMeF : abcès amibien foie.

Synonyme MeSH : Abscès hépatique amibien ; Amibiase du foie ; Amibiase hépatique ; Infection hépatique à Entamoeba.

Appartient aux [Méta-termes] : gastroentérologie ; hépatologie ; infectiologie ; parasitologie

Position du mot-clé dans l'abscès amibien foie

Vous pouvez consulter :

- toutes les ressources
- ou seulement les principales
- ou utiliser l'outil de recherche

05 mai 2011

courriel

Abcès amibien du foie

Copyright Memodata © 2005

(Menu général CISMeF) (haut de page)

© CHU de Rouen . Toute utilisation partielle ou totale de ce document doit mentionner la source.

Since March 2010, the bilingual version is used daily by 600 unique machines, mainly to query MeSH in their native language, as it was the case using the French MeSH Browser. Two hundred thirty people have already registered to access other T/O, mainly physicians, health students, librarians and translators. For the 11 T/O included in UMLS, we performed a contextual link to BioPortal (Noy *et al.* 2009).

4 Discussion

The Health Multi-Terminology Cross-lingual Portal (EHTOP) is used daily by various CISMeF academic partners in different French and European projects. In 2011, the Top 3 EHTOP targets are librarians, health professionals and health students to learn how to manipulate health terminologies (e.g. about rare disease with Orphanet thesaurus or anatomy with the FMA ontology) and to extract knowledge from it, in particular from hierarchies and relations (e.g. various siblings of a rare disease, symptoms of this rare disease or to obtain all the muscles of the forearm in one click). One important role of EHTOP is to improve the knowledge of health terms available in different T/O and in several languages, for example, to obtain the Latin translation of an anatomical term in English or French thanks to the FMA ontology. The fourth target is translators, who can access several health terminologies/ontologies in different languages. Combining EHTOP to Wordnet has improved both tools: allowing for EHTOP to obtain definitions and rich relations from Wordnet and for Wordnet

to obtain more health terms, including definitions and even richer relations (e.g. relations for one rare disease to the genes linked to it).

If French MeSH Browser was also heavily used (500 unique users per working day) mainly to access MeSH using queries in French, EHTOP allows access to MeSH in 10 other native languages. Via its Web services, EHTOP may also be used by several interactive applications. The standard response of the EHTOP Web Service is natively in the SKOS format. The other targeted users include the entire range of medical IT players (e.g. institutions, hospitals, software publishers, information portals) and, through them all those involved in the healthcare sector, in particular healthcare professionals and patients.

The EHTOP presented here has the main functionalities of any terminology server, except the extensive management of terminologies (e.g. adding a new hierarchy). To the best of our knowledge, EHTOP is the first of its kind to allow cross-lingual navigation. The main added value of EHTOP when compared to any UMLS browser (McCray et Razi, 1995) is the possibility to access the main health terminologies in French or the multi-lingual terminologies and classification coming from WHO, which are not (yet) included in the UMLS (e.g. ATC for drugs or ICPS for patient safety), as demonstrated in accessing ICD10 in five languages. Currently, the EHTOP is a necessary basic tool to index any document in a multi-terminology multilingual mode.

Other portals propose to search and navigate T/O such as NCBO Bioportal (Noy *et al.*, 2009) and

the EBI Ontology Lookup Service (Cote *et al*, 2006; Cote *et al*, 2008). Those tools are also very user-friendly and do not allow users to navigate through terms or search among synonyms in different languages. They are also not adapted to a daily use to index or to present the FMA to medical students.

In the near future, we have planned to integrate multi-lingual dictionaries (Wordnet) via an applet to the new multi-lingual health search engine Doc'CISMeF.

5 Conclusion

A health cross-lingual multi-terminology portal connected to a cross-lingual dictionary is a valuable tool to help to index and retrieve resources from a quality-controlled health gateway. It can also be very useful for teaching or performing audits in terminology management.

Acknowledgements

EHTOP was partially funded by PlaIR project, funded by FEDER (<http://www.plair.org>). The authors thank Richard Medeiros for his advice in the editing of this manuscript and the eight students of the INSA Rouen Engineering School that partially developed the multi-terminology portal.

References

- Burgun A, Denier P, Bodenreider O, Botti G, Delamare D, Pouliquen B, Oberlin P, Lévéque JM, Lukacs B, Kohler F, Fieschi M, Le Beux P. 1997. A Web terminology server using UMLS for the description of medical procedures. *J Am Med Inform Assoc* Sep-Oct; 4(5):356-63.
- Cote RG, Jones P, Apweiler R, Hermjakob H. 2006. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*. 28;7(1):97.
- Cote RG, Jones P, Martens L, Apweiler R, Hermjakob H. 2008. The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res*; 36:W372-6.
- Darmoni SJ, Leroy J-P, Baudic F, Douyère M, Piot J, Thirion B. 2000. CISMeF: a structured Health resource guide. *Methods Inf Med* 39(1):30-5.
- Darmoni SJ; Joubert M; Dahamna B; Delahousse J, Fieschi M. 2009. A French Health Multi-terminology Server. *AMIA symp*, 808.
- Fellbaum C. 2005. WordNet and wordnets. In: Brown, Keith et al. (eds.), *Encyclopedia of Language and Linguistics*, Second Edition, Oxford: Elsevier, 665-70.
- Lussier YA, Rothwell DJ, Côté RA. 1998. The SNOMED model: a knowledge source for the controlled terminology of the computerized patient record. *Methods Inf Med*, 37(2): 161-64.
- McCrory AT, Razi A. 1995. The UMLS Knowledge Source server. *Medinfo*, 144-7.
- Merabti T, Massari P, Joubert M, Sadou E, Lecroq T, Abdoune H, Rodrigues JM, Darmoni SJ. 2010. An Automated Approach to map a French terminology to UMLS. *Stud Health Technol Inform*, 1040-4.
- National Library of Medicine. 2010. Medical Subject Headings. Retrieved July 22, 2011, from <http://www.nlm.nih.gov/mesh/>
- Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey M-A, Chute CG and Musen MA. 2009. BioPortal: ontology and integrated data resources at the click of a mouse. *Nucleic Acids Research*; 37:W170-3.
- Sensegates. nd. Retrieved July 22, 2011, from <http://www.sensegates.com/>
- Thirion, B; Pereira, S; Névéol, A; Dahamna, B & Darmoni, SJ. 2007. French MeSH Brower: a cross-language tool to access MEDLINE/PubMed. *AMIA symp*, 1132.
- World Health Organization. nd. International Classification of Diseases, 10th revision. Retrieved July 22, 2011, from <http://www.who.int/classifications/icd/en/index.html>
- WordNet. A lexical database for english. nd. Retrieved July 22, 2011, from <http://wordnet.princeton.edu/>
- Zeng K, Bodenreider O. 2007. Integrating the UMLS into an RDF-Based Biomedical Knowledge Repository. *AMIA Symp*, 1170.

Revisiting a Brazilian WordNet

Valeria de Paiva

Rearden Commerce

Foster City, CA, USA

valeria.depaiva@gmail.com

Alexandre Rademaker

Applied Mathematics School, FGV

Rio de Janeiro, Brazil

arademaker@gmail.com

Abstract

Brazilian Portuguese needs a WordNet that is open access, downloadable and changeable, so that it can be improved by researchers, such as the community interested in automated deduction. This would be very valuable to linguists and computer scientist interested in representing knowledge obtained from texts. We discuss briefly why we want a Brazilian Portuguese WordNet and how we are going about getting one. These are only first steps, though, as our project is just starting.

1 Introduction

WordNet (Fellbaum, 1998) is an extremely valuable resource for research in Computational Linguistics and Natural Language Processing in general. WordNet has been used for a number of different purposes in information systems, including word sense disambiguation, information retrieval, automatic text classification, automatic text summarization, and dozens of other knowledge intensive projects.

But if there is still a lack of lexical resources for English, the problem is ten-fold more acute for other languages, which lack even easily accessible corpora and basic tools such as tokenizers, taggers and splitters. This lack of resources slows down considerably, almost stops completely any work on reasoning about knowledge obtained from language, our main goal.

We are starting a project at Fundação Getulio Vargas (FGV) in Brazil, where we want, in the long run, to use formal logical tools to reason about knowledge obtained from text in Portuguese. We are logicians, not linguists, so we want to minimize the amount of Computational Linguistics that we have to develop. Hence it

would have been very sensible to use a Brazilian WordNet, if we could have one. While we originally expected to be able to use some existing Brazilian WordNet, out of the box, it turns out that these are not available. There are some attempts.

There is the project WordNet.PT (Portuguese WordNet) from the “Centro de Linguitica da Universidade de Lisboa” headed by Palmira Marrafa. But this is available online only, no download available and, as far as we can see on their webpages, little development has happened recently to this project. The WordNet.PT version available online¹ has about 19000 lexical expressions, from different semantic fields. The fragment made available online includes expressions from subdomains such as art, clothing, geography, health, institutions, living entities and transportation, but no description of other domains and/or future releases of the database are discussed. The group has also a newer version of WordNet.PT called WordNet.PT_{global} (Marrafa et al., 2011) which pays attention to different varieties of Portuguese, like African variations in the language. But while this is very interesting for linguistic comparative research and useful for online queries², this smaller version of WordNet.PT is still not available for download and/or modifications and improvements.

Then there is also the MultiWordNet project and its Portuguese version MWN.PT, developed by Antônio Branco and colleagues at the NLX-Natural Language and Speech Group, of the University of Lisbon, Department of Informatics. According to their description³ MWN.PT the MultiWordnet of Portuguese (version 1) spans over 17,200 manually validated concepts/synsets, linked under the semantic relations of hyponymy and hypernymy. These concepts are made of over

¹<http://tinyurl.com/6p2vsy3>.

²<http://www.clul.ul.pt/wnglobal/>.

³<http://tinyurl.com/bum4mmh>.

21,000 word senses/word forms and 16,000 lemmas from both European and American variants of Portuguese. It includes the subontologies under the concepts of Person, Organization, Event, Location, and Art works, which are covered by the top ontology made of the Portuguese equivalents to all concepts in the 4 top layers of the English Princeton WordNet and to the 98 Base Concepts suggested by the Global Wordnet Association, and the 164 Core Base Concepts indicated by the EuroWordNet project. But again this wordnet is available online only and/or with a restrictive license that requires payment.

Thirdly there is a first version of a Brazilian Portuguese version of Wordnet developed by Bento Dias da Silva and collaborators (Dias-Da-Silva et al., 2000; Scarton and Aluisio, 2009). But this also cannot be downloaded, is not available online and is not being maintained in an open access basis, which is one of the strongest points of Princeton WordNet. Open access availability is one of the main reasons we would like to have our own version of WordNet-BR, which we are calling WN-BR. This is because we believe that resources like Wikipedia and WordNet need to be open and modifiable by others in order to improve over time.

Finally there is a whole batch of work on merging WordNet with Wikipedia categories and infoboxes, that tries to leverage the work already done by the Wikipedia volunteers. Amongst these we are particularly excited about YAGO/MENTA (de Melo and Weikum, 2010) (and YAGO2 for further work on temporal/spatial information), which we describe below. This kind of work fits in well with our goals of ultimately doing reasoning, in large scale, with knowledge obtained from text.

2 Global WordNet Grid

In this forum it is perhaps not necessary to recall that the Global WordNet Association aims at the development of wordnets for all languages of the world and to extend the existing wordnets to full coverage and all parts-of-speech. In 2006 the association launched a project to start building a completely free worldwide wordnet “grid”. This grid would be built around a shared set of concepts, and would be expressed in terms of the original Wordnet synsets and SUMO (Niles and Pease, 2001) terms. The idea of the grid is very appealing and the suggested procedure to create wordnets looks

sensible and feasible.

To recap the suggestion was to build the first version of the Grid around the set of 4689 “Common Base Concepts” and to make the Grid free, following the example of the Princeton WordNet. Now the Base Concepts are supposed to be the most important concepts in the various wordnets of different languages. The importance of the concepts was measured in terms of two main criteria: (1) A high position in the semantic hierarchy; (2) Having many relationships to other concepts. The procedure described as the “expand approach” seems to us viable: First translate the synsets in the Princeton WordNet to Portuguese, then take over the relations from Princeton and revise, adding the Portuguese terms that satisfy different relations. Then revise and revise and revise until we can guarantee the consistency of the taxonomy.

In somewhat more detail but still following the suggestions of the global wordnet grid, we think we can develop a core WordNet for Brazilian Portuguese by: first representing the 1024 core basic concepts by one or more synsets in Portuguese that are either equivalent or very closely associated to the original core concepts in the Princeton WordNet. Then adding Base Concepts that are important to Brazilian Portuguese, but not in the set of core basic concepts. (For that one could use lists of Portuguese words listed by frequency, and comparisons with other Wordnets for romance languages.) Next we need to check that this forms a closed and consistent hierarchy. Finally we should add further relations necessary to specify the semantics of the basic concepts.

While this procedure seems sensible and doable, despite being hard work, the existence of several versions of Wordnet, and the fact that the concepts uncovered as basic ones are not related to their synsets in WordNet 3.0 makes things more difficult. WordNet 3.0 has the huge advantage of disambiguated glosses, a real plus if the goal is semantics. But the mechanics of following the procedure sketched above turn out more complicated than expected.

Our solution is to use the mapping from WordNet 2.0 to WordNet 3.0 provided by (Daudé et al., 2000). The idea consists in, using the map, identify the WordNet 3.0 synsets equivalent to the 4689 WordNet 2.0 basic concepts listed in the EuroWordNet project (Stamou et al., 2002).

We plan to use the RDF version of WordNet 3.0 made freely available online⁴ by Mark van Assem and Jacco van Ossenbruggen. The RDF version has the benefit of better supporting our collaborative work, facilitating the maintenance in the same data structure of both English and Portuguese versions of the glosses and lexical forms, and, once loaded in a Triple Store, providing us with a working environment to add or remove relations, comments and so on. Unfortunately, there are at least two different versions of WordNet available in RDF. The RDF representation of WordNet 2.0 is described in <http://www.w3.org/TR/wordnet-rdf/> and seems to be used as reference for the newly minted RDF representation of WordNet 3.0. Nevertheless, we still have to investigate the differences between the WordNet 2.0/3.0 mapping used in the RDF representation of WordNet 3.0 and the mapping provided by (Daudé et al., 2000).

3 Sumo and WordNet-BR

The Global WordNet Grid approach has three aspects that seem to us worth considering. First there is the work on determining which synsets (corresponding to concepts) are most popular in several languages. This work was done in the EuroWordNet projects and it would be a shame not to use it. The emphasis on many languages should help filter out personal and cultural biases. Then there is the proposed stepping up schedule, which seems to us very attractive, as a way of helping to get the “grunt” work done. Finally and in some ways more importantly there is the connection with SUMO that we discuss now.

According to Wikipedia, the Suggested Upper Merged Ontology or SUMO is an upper ontology intended as a foundation ontology for a variety of computer information processing systems. It can be downloaded and used freely and it has been available and in development since 2000. A mapping from WordNet synsets to SUMO has also been defined and maintained for several versions of WordNet. Most importantly for us, SUMO is organized for interoperability of automated reasoning engines. In particular SUMO’s associated open source knowledge engineering environment, Sigma⁵ runs already in Vampire (Ganzinger et al., 1999) and Leo-II (Benzmueller and Paulson,

2010), for example. Projections of SUMO into description logics, automatically available, can be run in the OWL reasoners.

In the beginning of 2010 we started an informal project of discussing how logic and automated reasoning could have a bigger impact, if coupled with natural language processing and how it would be a great thing to translate some of the advances already made for English text understanding to Portuguese text understanding and reasoning.

Since one of us (Valeria de Paiva) had worked for almost nine years in Xerox PARC, in the systems developed by the Natural Language Theory and Technology (NLTT) group, particularly on the system Bridge, we requested an academic license to the XLE (Xerox Language Engine) to try to adapt the systems to Brazilian Portuguese. However, we are both logicians, our expertise lies at the end of the long pipeline of the system Bridge and we tried to recruit, still informally, people with more expertise on the language side of the project. But at that stage we did not have any formal backing, so despite some interesting offers, nothing much happened. Recently we have been granted formal backing, although still in small scale, and one of the opportunities that presented itself was to forestall the need for the creation of a Brazilian WordNet (or perhaps to help improve the creation of such), via the use of SUMO.

Wordnet is an important component of the XLE Unified Lexicon (UL (Crouch and King, 2005)), as the logical formulae created by the Abstract Knowledge Representation (AKR) component of the system are given meaning, in terms of Wordnet sysets. A previous version of the system used, instead of the Unified Lexicon, Cyc (Lenat, 1995) concepts as semantics. As discussed in (De Paiva et al., 2007) the sparseness of Cyc concepts was the main reason to move away from Cyc onto a version of the Bridge system based on the UL and WordNet. Since a WordNet-BR is not available, a workaround might be gotten via SUMO, if this were to be available in Portuguese. As a warming exercise we translated the basic concepts used for the basic concept descriptions in SUMO to Brazilian Portuguese⁶ and this is already available on the SourceForge repository for Sigma, SUMO’s knowledge engineering platform. This is not a substitute for a Brazilian Portuguese WordNet, but merely a stopping stone towards it.

⁴<http://semanticweb.cs.vu.nl/lod/wn30/>.

⁵<http://sigmakee.sourceforge.net/>.

⁶<http://tinyurl.com/bu874aq>

4 Scaling Up?

One of the ways we are considering of scaling up our proposal, from the five thousand concepts suggested in the grid page to the level that we think is necessary for our application goes via the work on YAGO (Suchanek et al., 2007) and, perhaps, YAGO2. The YAGO approach to information extraction for building a searchable, large-scale, highly accurate knowledge base of common facts goes via harvesting infoboxes and category names from Wikipedia for facts about individual entities. It reconciles these with the taxonomic backbone of WordNet in order to ensure that all entities have proper classes and the class system is consistent. The work in YAGO at the Max-PlanckInstitute has led de Melo and Weikum to work in MENTA (Multilingual Taxonomies from Wikipedia), which can be considered a multilingual version of WordNet. From this multilingual version (with 254 languages) we want to ‘project’ the component consisting of Portuguese synsets only. (The plan is to use the work in the Portuguese version of MENTA to complement, automatically, the five thousand concepts for WN-BR, obtained through manual translation. We believe that the MENTA (de Melo and Weikum, 2010) projection into Portuguese could give us a reasonable basis in terms of synsets in Portuguese to which we would like to compare the existing versions of Portuguese wordnets.) Since YAGO is already integrated with SUMO (De Melo et al., 2008), we hope to be able to maintain consistency of the database.

5 Conclusions

It is early days for our project and time will tell whether our decision to follow the global WordNet grid guidelines for seeding new wordnets will pay off or not, and if so, how well. We have now a master student interested in the project and more interested students are expected. One thing is clear to us, whatever kinds of resource we end up with, we hope to make them freely available in one of the numerous sites (SourceForge, GitHub, etc.) at our disposal nowadays. We are not aware of any such specialized lexicons available for Brazilian Portuguese and it is about time that we had them openly and freely accessible.

References

- C. Benzmueller and L.C. Paulson. 2010. Multimodal and intuitionistic logics in simple type theory. *Logic Journal of IGPL*, 18(6):881.
- D. Crouch and T.H. King. 2005. Unifying lexical resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbruecken, Germany.
- J. Daudé, L. Padró, and G. Rigau. 2000. Mapping wordnets using structural information. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 504–511. Association for Computational Linguistics. http://nlp.lsi.upc.edu/web/index.php?option=com_content&task=view&id=21&Itemid=57.
- G. de Melo and G. Weikum. 2010. Menta: inducing multilingual taxonomies from wikipedia. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1099–1108. ACM.
- G. De Melo, F. Suchanek, and A. Pease. 2008. Integrating yago into the suggested upper merged ontology. In *Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on*, volume 1, pages 190–193. IEEE.
- V. De Paiva, DG Bobrow, C. Condoravdi, R. Crouch, L. Karttunen, TH King, R. Nairn, and A. Zaenen. 2007. Textual inference logic: Take two. *Proceedings of the Workshop on Contexts and Ontologies, Representation and Reasoning*, page 27.
- B. C. Dias-Da-Silva, H. R. Moraes, M. F. Oliveira, R. Hasegawa, D. A. Amorim, C. Paschoalino, and A. C. Nascimento. 2000. Construção de um thesaurus eletrônico para o português do brasil. In *Processamento Computacional do Português escrito e falado (Propor)*, pages 1–10.
- C. Fellbaum. 1998. *WordNet: An electronic lexical database*. The MIT press.
- Harald Ganzinger, Alexandre Riazanov, and Andrei Voronkov. 1999. Vampire. In *Automated Deduction CADE-16*, volume 1632 of *Lecture Notes in Computer Science*, pages 674–674. Springer Berlin / Heidelberg.
- D.B. Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.
- Palmira Marrafa, Raquel Amaro, and Sara Mendes. 2011. Wordnet.pt global – extending wordnet.pt to portuguese varieties. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 70–74, Edinburgh, Scotland, July. Association for Computational Linguistics.

I. Niles and A. Pease. 2001. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 2–9. ACM.

Carolina Scarton and Sandra Aluisio. 2009. Herança automática das relações de hiperônimia para a wordnet.br. Technical Report NILC-TR-09-10, USP, São Carlos, SP, Brazil.

S. Stamou, K. Oflazer, K. Pala, D. Christoudoulakis, D. Cristea, D. Tufis, S. Koeva, G. Totkov, D. Dutoit, and M. Grigoriadou. 2002. Balkanet a multilingual semantic network for the balkan languages. In *Proceedings of the International Wordnet Conference, Mysore, India*, pages 21–25.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*, New York, NY, USA. ACM Press.

Scalar Properties of Emotion Verbs and Their Representation in WordNet

Christiane Fellbaum

Princeton University

Princeton, New Jersey, USA.

fellbaum@princeton.edu

Y. Yannick Mathieu

CNRS – Université Paris 7

Paris, France.

ymathieu@linguist.jussieu.fr

Abstract

We examine three groups of English verbs expressing the causation of emotions (*surprise*, *anger* and *fear*). Each verb in a given group denotes a different degree of intensity of the emotion. We search the Web for pairs of verbs co-occurring in specific lexical-semantic patterns like *if not...then...*, which have been shown to discriminate between the members of pairs of scalar adjectives in terms of their relative strength. Results show that the chosen verbs possess scalar qualities and confirm the prior assignment of the verbs into broad classes sharing an underlying emotion; finally, the Web data allow us to construct consistent scales with verbs ordered according to the intensity of the emotion. Furthermore, the Web data are compatible with scales constructed by human subjects, though across-subject variation is noted. We propose a way to represent the scalar aspects of verbs in WordNet that is consistent with current proposals for the integration of dimensional adjectives (Sheinman et al., submitted).

1 Introduction

The lexicon of emotions presents significant challenges for systematic investigation. Psychologists have identified a small number of basic emotions that are maximally distinct from one another and arguably have universal status, independent of cultural or linguistic diversity (e.g., Johnson-Laird and Oatley, 1989). Emotion verbs

have been classified semantically and syntactically for different languages (Belletti and Rizzi, 1988, Levin 1993, Mathieu 2005, *inter alia*), but there is little agreement among the different semantic analyses. We undertake a corpus-based examination of three groups of emotion verbs. Our goal is to empirically derive a semantic classification of English emotion verbs that

- (1) provides a novel, subtle and empirically grounded analysis of an important component of the English verb lexicon;
- (2) serves as the basis for appropriate representations in lexical resources serving NLP, such as WordNet;
- (3) has the potential to improve automatic text understanding by facilitating inferencing and the detection of lexically-based cohesion;
- (4) can be represented in WordNet in a way that is consistent with WordNet's structure.

2 Scalar Emotion Verbs

Psychologists and linguists have offered different classifications of emotions (e.g., Johnson-Laird and Oatley 1989, Martin and White 2005, Mathieu 2005, Ortony 1988). We examine three sub-groups of English emotion verbs (*frighten*, *surprise*, *anger*) based on Mathieu and Fellbaum's (2010) classification. Our goal is to rank the verbs in each group in terms of the intensity of the shared underlying emotion¹.

All classes consist entirely of so-called Experiencer psych verbs. As (1) shows, the transitive use of the verbs in this syntactically defined class requires a structural subject that refers to the

¹ Emotion verbs are not always easily distinguishable from cognition verbs. Thus, *shock* and *scandalize* denote events that can evoke both judgments and emotions. However, the possible overlap with cognition verbs does not affect our method or results.

Stimulus or Cause of the emotion, while the object expresses the Experiencer (Belletti and Rizzi 1988, Levin 1993). Thus, our class of *frighten* verbs includes the verbs *intimidate*, *scare*, *terrify* and *alarm* but not verbs like *fear* and *dread*, whose subject in a transitive construction denotes the Experiencer, as in (2):

- (1) *Lightning frightens/scares Mary*
CAUSE EXPERIENCER
- (2) *Mary fears/dreads lightning*
EXPERIENCER CAUSE

We depart from the observation that within each semantic class, the verbs express different degrees of intensity of the basic emotion that represents the class. For example, *terrify* seems intuitively to express a stronger degree of *frighten* than *scare*. Similarly, *anger* seems more intense than *annoy*, and *infuriate* in turn appears stronger than *anger*, though all three verbs refer to a feeling directed toward a real or perceived grievance. Hence these verbs represent different points on a scale, similar to gradable adjectives like *good-fine-superb* that express greater or smaller degrees of a common property (e.g., Bierwisch and Lang 1989, Kennedy 1999).

While gradability has been studied primarily for properties, which are typically expressed by adjectives, this semantic attribute clearly extends to concepts lexicalized by other parts of speech. Nouns like *fear* and *terror* and adjectives like *afraid* and *terrified* similarly differ in the strength of the emotion they express. Here we limit ourselves to the consideration of verbs.

3 Lexical-Semantic Patterns

To determine the relation among words with similar meanings, lexical-semantic patterns can be evoked. For example, Cruse (1986) noted that the patterns *Xs and other Ys* and *Ys such as Xs* establish that *Y* is a superordinate (more general) concept than *X*, and conversely, that *X* is a kind of *Y*. This is exemplified by phrases like *roses and other flowers* and *flowers such as roses*. Hearst (1992) and Snow et al. (2005), working with corpus data, extended this method and demonstrated its usefulness for detecting and encoding semantic relatedness. These authors have extracted lexical-semantic patterns to determine semantic relatedness for nouns; one goal is the extension of taxonomies and semantic networks like WordNet (Miller 1995, Fellbaum 1998).

Using the Web as a corpus, Sheinman and Tokunaga (2009) induce lexical-semantic patterns for gradable adjectives in order to determine their

relative position on a scale. They assume WordNet's organization of adjectives into "dumbbell" structures, with a pair of frequent and strongly associated "direct antonyms" such as *wet* and *dry* partitioning a scale into two. Each direct antonym is associated with a number of "semantically similar" adjectives, such as *drenched*, *soaked* at one side of the scale and *arid*, *parched* on the other (Miller 1998). From a given scale (e.g., "goodness"), Sheinman and Tokunaga (2009) select adjectives pairs, choosing first one of WordNet's direct antonyms (e.g., *good*) and one of its semantically similar adjectives from the same side of the scale (e.g. *great*). A search of the Web for occurrences of the expression *good*great* (where * is a lexically unspecified wildcard) yields lexical-semantic patterns consisting of single words or short phrases, such as (*perhaps*) *even* and *if not*². Inverting the order of the adjectives returns patterns like *if not...then* as in (3) and (4):

- (3) *good if not great*
- (4) *if not great, then god*

Sheinman and Tokunaga apply the patterns they induced to pairs of "semantically similar" adjectives and derive a relative ordering for them, which they subsequently evaluate against human judgments.

4 Ordering Verbs of Emotion on a Scale

Establishing the semantic relatedness of verbs is arguably much more difficult. Fellbaum (2002) and Chlovsky and Pantel (2004) propose the application of lexical-semantic patterns to detect and encode relations including and beyond those currently in WordNet. Chlovsky and Pantel propose an "strength" relation among verbs, but limit it to pairs.

We asked whether the patterns identified by Sheinman and Tokunaga (2009) for adjectives could be applied to emotion verbs, since many of these verbs resemble adjectives in exhibiting scalar properties and expressing different degrees of intensity of a common basic feeling.

As an initial proof of concept, we studied three broad classes that we judged to be clearly distinct and non-overlapping representatives of the emotion lexicon: *surprise*, *fear* and *annoyance*. Our

² Horn (1989) suggests that *even* serves to discriminate between more and less intense adjectives to the left and right of this word, respectively.

intuitive pre-classification assigned ten, five and nine members, respectively, to these groups:

Surprise verbs: *astonish, surprise, amaze, astound, strike, stun, floor, dumbfound, flabbergast, stupefy*

Frighten verbs: *intimidate, scare, frighten, alarm, terrify*

Annoy verbs: *irk, nettle, annoy, anger, exasperate, infuriate, enrage, incense*

It is important to note that we are merely interested in the relative ordering of the verbs on a scale and not making any claims as to the semantic distance between them. Quite possibly, some verbs are more similar to one another in terms of their intensity than others, and the distance among the verbs on the scale is not uniform, as seems to be generally the case with gradable adjectives³.

To arrive at a placement of the verbs on their respective scales, we applied Sheinman and Tokunaga's (2009) patterns. Like Sheinman and Tokunaga, we used the Web as a corpus, as its size constitutes an obvious advantage over balanced but relatively small corpora like the British National Corpus. We selected three patterns:

(P1) V_1 (*perhaps*) even V_2

(P2) V_1 , not to say V_2

(P3) If not V_1 then V_2

First, note that these patterns return verbs from the same syntactic class (Experiencer verbs with the Cause argument in subject position) on either side; the queries did not return hits with syntactically distinct emotion verbs, such as *she feared, even frightened him*.

Second, the directions of the patterns differ. In the case of the first and the second pattern, the more intense verb is found to the right of the phrase:

(5) annoy even infuriate

(6) anger, not to say incense

while in the third pattern, the verb expressing the more intense emotion is to the left:

(7) if not terrify then scare

5 Experimental Validation

Using the Google search engine, we manually searched all possible pairwise combinations, with both orders for each pair. Thus, for the five verbs

³ Just as in the case of scalar adjectives (Bierwisch and Lang 1989) the emotion verb scales are open-ended, independently of the lexical encoding of values on the scales. Thus, if *terrify* expresses the strongest form of *frighten*, this does not imply that *terrify* expresses the highest degree of fear a person may experience.

we assigned to the *fear* class (*alarm, frighten, intimidate, scare, terrify*) twenty pairs were searched for each patterns as shown Table 1. For each of the pairs, queries were written with each of the three patterns. The queries included the inflected forms of the verbs (past tense, past participle, third person singular present).

Table 1. Pairs of frighten verbs.

Verb ₁ -Verb ₂	Verb ₂ -Verb ₁
alarm-frighten	*frighten-alarm
*alarm-intimidate	*intimidate-alarm
alarm-scare	*scare-alarm
alarm-terrify	*terrify-alarm
intimidate-frighten	*frighten-intimidate
intimidate-scare	scare-intimidate
intimidate-terrify	*terrify-intimidate
frighten-terrify	*terrify-frighten
scare-frighten	*frighten-scare
scare-terrify	*terrify-scare

The results show for most pairs, one verb expresses a stronger degree of the same basic emotion than the other verb. The asterisks indicate the pairs for which either no example or only a single one were found. For the remaining cases, the number of hits ranged from dozens to hundreds. In the latter cases, we manually inspected at least a dozen returns to assure that the searches had returned valid examples. We manually identified and discarded some non-verb usages such as gerunds.

5.1 Results

Several points are worth noting. First, the *even* pattern was by far the most productive. This is unsurprising given that its meaning is roughly equivalent to the other two strong patterns while being lexically and syntactically much less complex and thus likely preferred by speakers.

Second, the majority of the hits contained the target not in its active verb form but as an adjectival past participle. Nevertheless, we believe that the participles encode the respective strength of the emotions in the same way as their underlying verbs and that our results are not impaired by the preponderance of adjectival rather than verbal forms. Quite possibly the patterns we used tend to favor adjectives over verbs as their collocates, as these patterns had been specifically identified by Sheinman and Tokunaga (2009) to occur with contrasting adjectives. Some of the target verbs have homographic nouns (*alarm*,

scare) and a search for the base form of the verb often returned hits with the nouns.

Below we illustrate our findings with representative examples from the group of five verbs expressing fear-causing events. We present examples for each of the pairs we found from the set of twenty possible pairs listed in Table 1. The data are similar to those for the two other verb groups that we examined (*surprise, annoy*) but these will not be presented and discussed here in detail.

(8) *Truly, a cold chill gripped me, my heart rate increased, and I became **alarmed, even frightened**.*

(9) *She looked **alarmed, even scared** by our festive look.*

(10) *Now, when money is tight, the Government seeks to claw back the entitlements and does so in a way that **alarms, even terrifies***

(11) *The ticking clock **intimidates us, even frightens us***

(12) *I am still dealing with some level of doubt and fear, the assignment **intimidates and even scares me** a little.*

(13) *However, for some people, anger is a constant companion that defines their personality and intimidates - even terrifies - their loved ones*

(14) *The prospect of change and evolution frightens, even terrifies, many people around the world*

(15) *For the first time since she had crossed him Annabeth looked scared, even frightened.*

(16) *See something in McCain that others don't; something that scares - even terrifies - many of those who know McCain on a personal basis.*

(17) *When she first became an agent on his team she felt scared even intimidated by them*

5.2 Outliers

Scare and *intimidate* were found on both sides of the patterns, as exemplified in (12) and (17). Although our search returned several hundred examples where *scare* expressed a stronger emotion than *intimidate*, only six examples with the inverse order, such as (17), were found, all with the *even* pattern. A possible explanation for the fact that *scare* and *intimidate* do not seem to exhibit a clear asymmetry might be that their meanings differ not just with respect to the intensity of the same underlying emotion but that there is an added subtle meaning difference. As a result, the two verbs, though related semantically, might not fall neatly on the same linear scale of intensity. This explanation receives some support from the

results for the pair *alarm-intimidate*, which did not co-occur in either order, with the exception of one single example. This case, too, suggests that the verbs, though similar, are not members of the same broader *frighten* class. Given the overall results, *intimidate* seems to be the outlier. Moreover, we did find numerous examples where these verbs occur in a coordinate structure:

(18) *Being ignorant and illiterate, he was **alarmed and intimidated** by the fear that his land would be taken from him by Fletcher*

(19) *If you receive a "reservation of rights" letter from the insurance company, don't be **alarmed or intimidated***

And and *or* do not discriminate among scale mates; thus, phrases like *good and superb* or *speedy or fast* seem odd. These examples indicate further that the paired verbs here are semantically similar but not necessarily scale mates with different degrees of intensity. In other cases, the verbs occurred in a "list" pattern:

(20) *Intimidation related to prejudice and discrimination may include conduct **which annoys, threatens, intimidates, alarms, or puts a person in fear of...***

(21) *If someone else's behavior makes you feel **alarmed, intimidated, embarrassed, or annoyed**, there are many resources available*

Such "listing" does not reveal an asymmetry in the strength of the emotion, though it indicates that the emotions are similar and can be evoked by a common Cause or Stimulus.

6 Establishing a Scale

The data summed up in Table 1 show asymmetries among the members of the verb pairs, as suggested by the lexical-semantic patterns. Considering our data in the light of the two different kinds of patterns—two in which the more intense verb occurs on the right and one where it occurs to the left—we can construct a scale.

For example, we can place *scare* (less intense) to the left of *terrify* (more intense) on the basis of examples like the following:

(22) *Are there other women out there with SA who are scared (even terrified) of becoming pregnant and being responsible for a new life*

(23) *He is terrified of being on stage - if not terrified then scared, worried, mildly anxious, you get the drift.*

And we can place *scare* to right of *alarm*, since it seems stronger:

(24) *However, your solution of involving the private sector is one that the public should be very alarmed, if not scared*

Based on such examples we construct a linear scale:

Intimidate>alarm>scare>frighten>terrify

For the class of surprise verbs, which comprised ten verbs, queries involving the verbs *amaze*, *astonish* and *surprise* returned large numbers of hits where these verbs appear on both sides of the patterns. We interpret this to suggest that speakers use these verbs to refer to very similar, if not identical, degrees of surprise. Moreover, these three verbs also pattern identically with respect to the other verbs on the scale (i.e., *amaze*, *astonish* and *surprise* were all consistently weaker than *stun* and *dumbfound*, for example).

6.1 Cross-Class Pairs

We applied the patterns with only a single verb from our classes on either side. In some cases, these searches returned hits where the second verb did not come from the same group of emotion verbs preclassified by Mathieu and Fellbaum (2010).

(25) *offended and alarmed, if not threatened by the song*

(26) *The Philadelphia Fringe Festival is growing, but many intelligent theater-lovers are puzzled and even intimidated by it*

Such data may suggest that the assignment of verbs to their classes needs to be reconsidered. However, it seems more likely that the verbs belong to different but semantically similar classes that may have fuzzy boundaries; the lexical-semantic patterns may “leak” across these related classes. However, we did not find examples where the verbs on either side of the pattern were in completely unrelated classes. For example, no hits of the kind **this surprised, not to say exhausted/reminded me* were returned.

7 Human judgments of scalar emotion verbs

To compare the automatically derived scales against human judgments, we collected data from ten Princeton University students, all of them native speakers. They were given a booklet with the three groups of verbs expressing causation of fear, surprise, and anger. The verbs in each group were presented in random order, and the students were asked to construct

scales based on the intensity of the emotion expressed by all members of a scale. The scale *like>love>adore* served as an illustrative example. Options to construct multiple scales for verbs in a given group and to place several verbs on the same point on a scale were explicitly allowed for and exemplified.

7.1 Results of Human Judgments

Tables 2-4 show how often the judges placed a given verb on a position on the scale. Interestingly, the students—who were not given the lexical-semantic patterns-- stated that they found the task very difficult.

The tables reveal the relative strength of the verbs on a scale, in that the number of placements for a given verb increases or decreases linearly (e.g., *irk*) or shows a local peak (e.g., *amaze*). While not as clear-cut as the results obtained automatically via the AdjScale method, they are clearly consistent with the corpus-derived data. The perceived difficulty of the task may be blamed in part; introspection about subtle meaning differences is difficult (Fellbaum, Grabowski and Landes, 1997). As one would expect, the results are noisier for the larger verb groups of *anger* and *surprise* verbs.

Table 2. Human judgments of scalar verbs of causing fear.

Verb	Pos.	#1	#2	#3	#4	#5
<i>intimidate</i>		5		2	2	
<i>alarm</i>		4	4	1	1	
<i>scare</i>		1	4	4	1	
<i>frighten</i>			4	2	4	
<i>terrify</i>				1	1	8

The data show that most participants judged *terrify* to be the verb that expresses the strongest degree of causation of fear while *intimidate* was judged to be the weakest verb in this scale. Two participants judged *scare* and *frighten* to be equally strong, placing them both in position Two. One participant stated that *intimidate* does not belong on the scale of verbs denoting causation of fear. This conforms nicely to the results of our automatic analysis, which pointed to *intimidate* as an outlier.

Table 3. Human judgments of scalar verbs of causing anger.

Verb	Pos.	1	2	3	4	5	6	7	8	9
<i>nettle</i>		6	3		1					
<i>irk</i>		5	2	3						
<i>annoy</i>		4	5	1	1					
<i>irritate</i>			3	4	2					

Verb	Pos.	1	2	3	4	5	6	7	8	9
<i>exasperate</i>				3	3	2			2	
<i>anger</i>			1	1	2	4	1			
<i>incense</i>				1		4	3	2	1	
<i>enrage</i>					2	2	3	2	1	
<i>infuriate</i>					1	2	2	1	1	2

Irk and *nettle* were judged to be the weakest forms of angering, while *incense*, *enrage* and *infuriate* were judged to be the strongest; the ratings reveal that these verbs were not very clearly distinguished in terms of their intensity.

Table 4. Human judgments of scalar verbs of causing surprise

Verb	Pos.	1	2	3	4	5	6	7	8	9	10
<i>strike</i>		6	3	1	1						
<i>surprise</i>		5	6								
<i>amaze</i>		1	2	5	1						1
<i>dumbfound</i>			1		2	1	1	2	2		1
<i>astound</i>				1	3	3	1		1	1	
<i>astonish</i>				3	3	2		2			
<i>flabbergast</i>				1	1		2	2	3		1
<i>stun</i>				3	5	1		1			
<i>floor</i>					1	1	5	1	1		
<i>stupefy</i>					1			3	1	3	

These verbs seem to be the most difficult to distinguish in terms of their intensity. While *strike*, *surprise* and *amaze* were clearly judged to be the least strong verbs, the participants did not assign clear intensity ratings to *dumbfound*, *astound*, and *flabbergast* with respect to other verbs in the group.

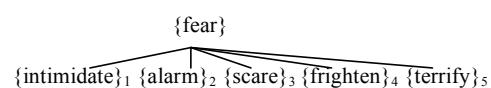
The collection of human judgments concerning the relative order of experiencer verbs based on intensity yielded less clear-cut results than the corpus searches. But importantly, both sets of the results are consistent with each other.

8 Representing Verbs with Scalar Properties in WordNet

Verbs of emotion, like all other verbs, are currently encoded in WordNet in terms of troponymy (the super-/subordinate "manner" relation), entailment and antonymy. This representation does not make reference to different degrees of intensity, though troponyms of a common superordinate are likely to be verbs expressing not only more specific but also stronger forms of the emotion. For example, the synset *{hate, detest}* currently has two immediate subordinates, *{abhor, loathe, abominate, execrate}* and *{contemn,*

despise, scorn, disdain}. Its antonym, *{love}* has three direct troponym synsets, *{care for, hold dear, cherish, treasure}*, *{dote}* and *{adore}*. Clearly, the assignment of these verbs to synsets seems somewhat arbitrary and does not try to reflect subtle differences in the intensity of the emotion either among synset mates nor across synsets. Introducing scales that encode the relative strengths of a shared emotion among different verbs would be desirable and useful for various applications (see "Future Work").

Sheinman et al. (submitted) make a specific proposal to represent scalar dimensional adjectives (*big, large, huge, gigantic, enormous,...*) in WordNet in a way that reflects their relative strength. Briefly, WordNet's current adjective dumbbells are maintained but enhanced with links to the noun synsets that refer to the related attribute (thus, *big, large, etc.* are linked to *size*). Additionally, those dumbbell members for which corpus data indicate a partial ordering of strength are represented on an appropriate linear scale. We propose to adopt the scalar representation for verbs of emotions as well. Each synset is marked by an integer referring to its position on the scale. The members of a scale are all linked to a noun that expresses the common emotion. An example is shown in Figure 1:



The representation of other verb-verb relations (entailment) will remain unchanged.

9 Limitations

Our work is preliminary; many more verbs and patterns need to be investigated. The number of hits returned by our corpus queries for a given pair depends on the frequency of the verbs in the query, for which counts in the British National Corpus serve as a guideline. Thus, for *surprise* verbs, queries including *flabbergast* returned far fewer hits than queries for *astonish* and *amaze*. Moreover, the limitations of even a very large corpus like the Web allow for the possibility that not finding any corpus examples of a given pattern does not mean that this pattern is inadmissible.

10 Conclusion and Future Work

Encoding the relative ordering of verbs on a scale reflects their meaning more clearly than is currently the case in WordNet. The degree of intensity of the emotion shared by all scale mates is perhaps the verbs' most important semantic component, since it distinguishes the scale mates from one another.

Some research in sentiment analysis classifies emotion terms that characterize a speaker's or writer's attitude towards a particular issue in a binary fashion, distinguishing only "positive" from "negative" emotions (e.g., Turney 2002, Yu and Hatzivassiloglou 2003). A binary classification seems overly coarse-grained, given the number of distinct lexemes that denote varying strength of a common emotion. Rather than assigning verbs to one of two opposed poles, our work attempts to determine more fine-grained distinctions among semantically similar verbs of emotions. Representing the scalar properties of some verbs enables inferencing: Knowing that a verb v_1 is placed to the right of another verb v_2 on the same scale of intensity (i.e., that v_1 is stronger or more intense than v_2) means that if $X v_1 \rightarrow Y$ then X necessarily $v_2 \rightarrow Y$. Thus, if Slovakia's knocking out England in the World Cup *dumbfounded* John, it necessarily *surprised* him. (The entailment is, of course, unidirectional.)

The results reported here are intended to serve as a proof of concept, and we are currently extending the approach to other areas of the lexicon and applying additional patterns to the automated searches so that the results can be scaled up and be considered more robust.

Finally, a crosslinguistic examination of scalar properties of emotion verbs might reveal interesting differences with respect to lexicalization. Not all languages are likely to lexically encode the same degrees of a given emotion. While the work reported here cannot reliably reveal crosslingual matches, it can show how densely the semantic space around a given emotion is lexically labeled and distinguished.

Acknowledgments

Fellbaum's work is supported by grant number CNS 0855157 from the U.S. National Science Foundation and the Tim Gill Foundation.

References

- Belletti A., Rizzi L.: Psych-Verbs and θ -theory. *Natural Language and Linguistic Theory* 6: 291-352. (1988).
- Bierwisch M., Lang E., Eds.: Dimensional adjectives: grammatical structure and conceptual interpretation. Springer, Berlin (1989).
- Chlovsky, Timothy and Patrick Pantel. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. Conference on Empirical Methods in Natural Language Processing (EMNLP), 33-40.
- Cruse D. A.: Lexical semantics. Cambridge University Press, Cambridge, UK (1986).
- Fellbaum C. The English Verb Lexicon as a Semantic Net. *International Journal of Lexicography*, 3/4, 278-301 (1990).
- Fellbaum C. Co-Occurrence and Antonymy. *International Journal of Lexicography*, 8/4, 281-303 (1995).
- Fellbaum C., Grabowski J., Landes S.: Analysis of a Hand-Tagging Task. In: ACL/Siglex Workshop (1997).
- Fellbaum C., Ed.: WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA (1998).
- Fellbaum, C. 2002. Parallel Hierarchies in the Verb Lexicon. In: Simov, K. (Ed.), Proceedings of the Ontolex Workshop on Ontologies and Lexical Knowledge Bases, 27-31. Paris, France: Elra.
- Hearst, Marti. (1992). Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th conference on Computational Linguistics (COLING), 539-545.
- Horn L. R.: A Natural History of Negation, University of Chicago Press, Chicago (1989).
- Johnson-Laird P. N., Oatley K. J.: The language of emotions: an analysis of a semantic field. *Cognition and Emotion* 3:81-123 (1989).
- Kennedy C.: Projecting the adjective, Garland Publishing, New York & London (1999).
- Levin B.: English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago (1993).
- Martin J. R., White P. R. R.: The language of evaluation, Palgrave Macmillan (2005).
- Mathieu Y. Y.: A Computational Semantic Lexicon of French Verbs of Emotion, Shanahan, G., Qu, Y., Wiebe, J. (eds.): *Computing Attitude and Affect in Text*, Springer, Dordrecht (2005).
- Mathieu Y. Y., Fellbaum C.: Verbs of Emotion in French and English. In: The 5th Global WordNet Conference, 70-75 (2010).
- Mathieu, Y.Y., Fellbaum, C. (2011). A corpus-based examination of scalar verbs of emotion. Paper presented at the Conference on Cognition, Emotion and Communication. Nicosia, Cyprus.
- Miller G. A.: Wordnet: A lexical database for English, *Communications of the ACM* 38: 39-41 (1995).
- Miller K. J.: Modifiers in WordNet. Fellbaum, C. (Ed.): WordNet. An electronic Lexical Database: 47-67 (1998).
- Ortony A., Clore G. L., Collins A.: The Cognitive Structure of Emotions. Cambridge University press, Cambridge (1988).
- Sheinman V., Tokunaga T.: AdjScale: Visualizing differences between adjectives for language learners. IEICE Transaction of Information and Systems, Vol.E92-D, No.8: 1542-1550 (2009).
- Sheinman, Fellbaum, Julien, Schulam, Tokunaga (submitted). Large, Huge, or Gigantic? Identifying and encoding intensity relation among adjectives in WordNet. *Language*

- guage Resources and Evaluation: Special issue on wordnets and relations.*
- Snow, Rion, Daniel Jurafsky and Andrew Ng. (2005). Learning syntactic patterns for automatic hypernym discovery. *Advances in neural information processing systems* 17: 1297-1304.
- Turney P.: Thumps up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: The 40th Annual Meeting of ACL, Philadelphia, Pennsylvania (2002).
- Yu H., and Hatzivassiloglou V.: Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In: *EMNLP2003*: 129-136 (2003).

sloWNet 3.0: development, extension and cleaning

Darja Fišer

University of Ljubljana
Ljubljana, Slovenia

darja.fiser@ff.uni-lj.si jernej.novakl@uni-mb.si

Jernej Novak

University of Maribor
Maribor, Slovenia

Tomaž Erjavec

Jožef Stefan Institute
Ljubljana, Slovenia

tomaz.erjavec@ijs.si

Abstract

In this paper we present the development, extension and cleaning of Slovene wordnet by reusing existing language resources. The initial induction of synsets and the subsequent extension of sloWNet are based on multilingual resources and were performed automatically. The cleaning of the developed lexicon, on the other hand, is based on a monolingual reference corpus and requires manual validation. Manual work is performed in sloWTool, a new browser, editor and visualizer of wordnet content. The developed wordnet and editor are freely available under the Creative Commons licence.

1 Introduction

In the past five years much effort has been invested to close the gap in Slovene language resources which had still been lacking in the lexico-semantic layer. Semantic lexica and semantically annotated corpora are a prerequisite for practically any task that involves semantically enhanced processing of natural language. In addition, they are also extremely useful in applied linguistics, such as lexicography and language pedagogy, as well as in corpus linguistics for the study of sense frequency and co-occurrence.

While the development of the resources is still on-going and a lot of work on further extensions and refinements still needs to be done in the future, we have reached a relatively stable version of the resources which are large and precise enough to be useful in practical applications. The Slovene wordnet called sloWNet has been used to mine definitions from corpora (Fišer et al. 2010) improve the results of machine translation at lexical level (Fišer and Vintar 2010), to automatically detect semantic shifts in translated

texts (Vintar 2011), and as a seed dictionary for building context vectors to extract bilingual lexicons from large comparable corpora (Ljubešić and Fišer, submitted).

The aim of this paper is to present the latest developments connected with the Slovene wordnet and is organized as follows: in the next section we summarize the development stages of sloWNet and report on the content of the latest version. In Section 3 we present the corpus that was annotated with wordnet synsets and the refinements that were made in order to achieve greater consistency. In Section 4 we describe the tool we developed for browsing, editing and visualizing wordnet content, and in Section 5 conclude with some final remarks and plans for future work.

2 Automatic development of sloWNet

Slovene wordnet is based on Princeton WordNet (Fellbaum 1998) and was built automatically in three stages, each using a different approach according to the resources used for extracting the relevant lexico-semantic information. The first and most straightforward approach relied on the Serbian wordnet (Krstev et al. 2004) where the literals were translated into Slovene utilizing a traditional digitized bilingual Slovene-Serbian dictionary (Erjavec and Fišer 2006). This simple approach lacked automatic disambiguation of polysemous dictionary entries and therefore required a lot of manual cleaning. This was improved in the second approach which was able to assign the correct wordnet sense to a Slovene equivalent by disambiguating it with a word-aligned parallel multilingual corpus and already existing wordnets for several languages (Fišer 2007). The main contribution of the third and final approach was the extraction of a large number of monosemous specialized vocabulary and

multi-word expressions from Wikipedia and its related resources (Fišer and Sagot 2008). After merging the results of these three approaches sloWNet contained about 17,000 literal which belonged to roughly 20,000 synsets.

3 Further extension of sloWNet

The next major step in the development of sloWNet 3.0 is the recent large-scale automatic extension in which we combined all the resources from the previous steps in order to exploit the available resources to their full potential and thereby improve coverage of sloWNet without compromising its quality. First, a model was trained on the existing elements in sloWNet, and a maximum entropy classifier was used to determine appropriate senses of translation candidates extracted from the heterogeneous resources described above (see Sagot and Fišer 2012).

The extended sloWNet now has 114% more synsets than before, while the number of (*literal, synset*) pairs has increased from 24,081 to 82,721, which is 244% of its initial size. An analysis of the sources that contributed to creation of the current synsets in sloWNet shows that for adjectives, the classifier provided as much as 95% of the synsets. Similarly, 88% of the adverbial synsets were populated by the classifier, while the rest were added by hand during manual revision of the created wordnet (see Secton 4). Verbs behave in a very similar fashion with 85% of them originating from the classifier while the rest originated in the previous development step from the dictionary, manual revision and the corpus. This is understandable because these parts of speech had been handled less successfully in our previous wordnet development approaches. But a somewhat different distribution can be seen among nominal synsets that had dominated the Slovene wordnet all along. Of these, 36% had already been created before from Wikipedia, dictionary and the corpus or were added manually, so that only 64% were contributed in the recent extension process by the classifier.

As Table 1 shows, the current version of Slovene wordnet contains 36% of all the synsets in Princeton WordNet. Nouns are still by far the most frequent, representing more than 70% of all synsets. sloWNet contains all synsets from the Base Concept Sets but also has a lot of specialized vocabulary as 66% of all the literals in it are monosemous. The extended sloWNet also contains a lot of multi-word expressions and proper names, which are both mostly nominal. A com-

parison of the average number of literals per synset and average level of polysemy between sloWNet and PWN is interesting because it can indicate how accurate the automatic population of Slovene synsets was. While average synset length is comparable to PWN, the total average polysemy (2.07 vs. 1.51) and the average polysemy excluding monosemous words (4.19 vs. 3.39) show that Slovene wordnet contains noise that will have to be filtered out in the future (see Section 4).

The fact that sloWNet is noisy due to the automatic construction process is further indicated by the number of literals in the longest synsets which are, at first glance, quite similar to PWN (see Table 2) but a more careful analysis shows that even though these synsets contain several synonyms, not all of them are correct and should therefore be filtered out in the future.

This is even more obvious with the most polysemous literals in sloWNet that are clearly very noisy. The most important source of such errors was the inadequate sense assignment for the most frequent words in the language, such as the verb *to be*, the noun *person*, the adjective *big* and the adverb *very*.

While Princeton WordNet contains glosses for all its 117,658 synsets, sloWNet currently contains only 3,178 definitions for nominal synsets that were extracted automatically from Wikipedia articles. 32,881 PWN synsets are also equipped with at least one usage example which is only the case for the 517 sloWNet nominal synsets that were annotated in the corpus. A focused attempt to providing additional definition and example sentences is planned in the near future.

As additional information, useful in for many tasks, many wordnet synsets have domain labels which are further organized into a domain hierarchy (Bentivogli et al. 2004). Analysis of the domains in the created Slovene wordnet shows that they are much better represented. 46% of all the synsets in PWN that belong to one of the domains exist in sloWNet as well. Of all 161 domains that are present in PWN, only 4 of them are missing entirely, all of them belonging to the Sports domain hierarchy: Rugby, Soccer, Sub and Volleyball. Just like in PWN, the most frequent domain is Factotum and the following three most frequent ones are represented in the same order in both wordnets. There are also many similarities among the ten most frequent domains in the two wordnets (see Table 3).

no. of synsets			no. of literals			no. of (synset, literal) pairs		
	PWN3.0	sloWNet3.0		PWN3.0	sloWNet3.0		PWN3.0	sloWNet3.0
Adj	18,156	6,218	Adj	21,538	5,108	Adj	30,004	12,438
Adv	3,621	453	Adv	4,481	514	Adv	5,580	847
N	82,114	30,911	N	119,034	30,319	N	146,345	55,383
V	13,767	5,337	V	11,531	3,840	V	25,047	14,053
total:	117,658	42,919	total:	156,584	39,781	total:	206,976	82,721
BCS1	1,220	1,220	monos.	130,208	26,339	avg. syn. length	1.76	1.92
BCS2	2,213	2,213	mwe	64,383	9,050	avg. polys.-all	1.51	2.07
BCS3	1,238	1,238	proper n.	35,002	2,946	avg. polys.-poly	3.39	4.19
total:	4,671	4,671	non-letter lit.	178	32			

Table 1: A comparison of Princeton WordNet 3.0 and sloWNet 3.0

longest synsets		
POS	PWN 3.0	sloWNet 3.0
Adj	23 (02074929-a)	23 (00148078-a)
Adv	10 (00048739-b)	14 (00004722-b)
N	28 (05559256-n)	20 (05921123-n)
V	25 (01426397-v)	24 (00933821-v)
most polysemous literals		
POS	PWN 3.0	sloWNet 3.0
Adj	27 (heavy)	47 (velik~big)
Adv	13 (well)	13 (zelo~very)
N	33 (head)	70 (oseba~person)
V	59 (break)	757 (biti~to be)

Table 2: A comparison of longest synsets and most polysemous literals in PWN 3.0 and sloWNet 3.0

PWN 3.0	Synsets	sloWNet 3.0	Synsets
<i>Factotum</i>	19,454	<i>Factotum</i>	9,701
<i>Zoology</i>	6,270	<i>Zoology</i>	3,345
<i>Botany</i>	5,998	<i>Botany</i>	2,716
<i>Biology</i>	3,004	<i>Biology</i>	1,512
<i>Gastronomy</i>	2,183	Person	793
<i>Chemistry</i>	2,011	Admin.	790
<i>Medicine</i>	1,999	Chemistry	656
<i>Admin.</i>	1,909	Medicine	625
<i>Anatomy</i>	1,768	Building_ind.	575
<i>Person</i>	1,600	Gastronomy	525
Total	77,701	total	33,126

Table 3: A comparison of synsets belonging to domains in PWN 3.0 and sloWNet 3.0

4 Cleaning of sloWNet synsets

Even before the large-scale automatic extension, sloWNet has undergone two cycles of partial

manual revision; the goal of the first revision was to manually check, correct and add the missing translations for all synsets belonging to the Base Concept Sets (about 5,000) while the second revision was performed in parallel with semantic annotation of the corpus (see Fišer and Erjavec 2010). At that time, all the synsets containing the nouns (about 1,000, not all of which were finally assigned to words in the corpus) which were selected for annotation in the corpus were checked and corrected as necessary and the missing senses of those nouns were also added if they were found in the corpus.

But as the analysis of the automatically extended wordnet in the previous section shows, a more comprehensive cleaning of the resource is required. We have developed a 2-step procedure where we first automatically detected those literals that are most likely outliers given the synsets they appear in, which were then presented to wordnet editors for manual validation.

The automatic detection of noisy literals is based on distributional methods and aims to identify the most obvious errors in synsets that occurred due to errors in word-alignment of parallel corpora (e.g. misaligned elements of multi-word expressions) and inappropriate word-sense disambiguation of homonymous words (e.g. assigning a valid translation of one sense of a homonymous source word to all its senses). We started from a (noisy) list of synonym candidates and ranked them according to the similarity of contexts they appear in FidaPLUS reference corpus (Arhar and Gorjanc 2007). The ranking relied on a simple hypothesis that literal-synset pairs tend to co-occur in corpora with other lexemes that are semantically related, as made explicit by relations between synsets in wordnet (see Sagot and Fišer, submitted).

So far, the procedure has been applied on nominal synsets only because they represent the majority (67%) of all synset-literal pairs in the latest version of sloWNet. Of all nominal synset-literal pairs in sloWNet, 37,356 (67%) were attested in the reference corpus. We then empirically set a threshold that defines the minimum score under which a (literal, synset) pair is considered as a candidate outlier. The overall error rate in the extended sloWNet has been evaluated at 15%, which means that around 13,000 incorrect (literal, synset) pairs were introduced with the automatic extension. We have therefore chosen a threshold for candidate outliers of the same order of magnitude, generating 12,578 candidate outliers. They are not deleted from wordnet automatically because some candidates could not be ranked reliably due to lacking distributional information in the corpus, which is why they are presented to wordnet editors who decide whether they are true outliers or not. Despite the manual effort required in the cleaning of the created wordnet, the approach is still valuable because instead of having to check all (literal, synset)

pairs in sloWNet, the editors now check only about a third of them, saving them a significant amount of time and effort. With the experience gained during manual validation of the candidate outliers we plan to further refine their automatic detection as well as to extend the approach to other parts of speech.

The detected candidate outliers are currently being manually validated in slowTool, an all-in-one browser, editor and visualizer for wordnets we developed for this and many other tasks (Fišer and Novak 2011). An example of this procedure is illustrated in Figure 1 where the polysemous English word *organ* has two possible translations in Slovene: *organ* for the body part sense, and *orgle* for the instrument sense. In the example shown, *orgle* is the correct literal for the synset but *organ* is not, which was correctly detected as an outlier candidate and therefore has to be manually deleted from the synset by clicking on the delete button right of the highlighted literal.

Figure 1: Manual editing of synsets in slowTool.

5 Conclusions

In this paper we presented a set of automatic approaches we used to develop Slovene wordnet by recycling already available language resources. The analysis of the created sloWNet showed that filtering of inappropriate synset elements was required in order to reduce the noise in the resource, which is being done in a 2-step way by first automatically detecting candidate outliers based on distributional methods and then manual validation of the noisy candidates. Manual work is being performed in sloWTool, the 3-in-1 tool for wordnet browsing editing and visualisation. Both sloWNet and slowTool are freely available for research under the Creative Commons license at <http://nl.ijz.si/slownet/>.

In the future we plan to refine the approach to removing noisy synset candidates so that it can handle more subtle polysemy as well and to extend it to other parts of speech. We also wish to focus on refinements and extensions of the sloWNet-annotated jos100k corpus (Erjavec et al. 2010) so that it can serve as a training set for automatic word-sense disambiguation. In addition, we are looking into possibilities to further develop sloWTool to allow assigning wordnet senses to words in the corpus.

References

- Arhar, Špela and Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa (The FidaPLUS corpus: a new generation of the Slovene reference corpus). *Jezik in slovstvo*, 52(2).
- Luisa Bentivogli, Pamela Forner, Bernardo Magnini and Emanuele Pianta. 2004. Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. *Proc. of the Workshop on Multilingual Linguistic Resources*, COLING'04, Geneva, Switzerland, August 28, 2004, pp. 101-108.
- Tomaž Erjavec and Darja Fišer. 2006. Building the Slovene Wordnet: first steps, first problems. *Proc. of the Third International WordNet Conference (GWC'06)*, Jeju Island, Korea, January 22-26, 2006.
- Tomaž Erjavec, Darja Fišer, Simon Krek and Nina Ledinek. 2010. The JOS linguistically tagged corpus of Slovene. *Proc. of 7th International Conference on Language Resources and Evaluation (LREC'10)*, Malta, May 17-23.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Darja Fišer. 2007. Leveraging parallel corpora and existing wordnets for automatic construction of the Slovene wordnet. *Proc. of the 3rd Language & Technology Conference*, October 5-7, 2007, Poznań, Poland, pp. 162-166.
- Darja Fišer and Benoît Sagot. 2008. Combining multiple resources to build reliable wordnets. *Text, Speech and Dialogue (LNCS 2546)*. Berlin; Heidelberg: Springer, 2008 pp. 61-68.
- Darja Fišer and Špela Vintar. 2010. Uporaba wordneta za boljše razdvoumljanje pri strojnem prevajaju. *Proc. of the 13th International Multi-conference Information Society (IS'10)*.
- Darja Fišer, Senja Pollak and Špela Vintar. 2010. Learning to mine definitions from Slovene structured and unstructured knowledge-rich resources. *Proc. of 7th International Conference on Language Resources and Evaluation (LREC'10)*, Malta, May 17-23 2010.
- Darja Fišer and Jernej Novak. 2011. Visualizing sloWNet. *Proc. of Electronic lexicography in the 21st century: new applications for new users (eLex'11)*, Bled, 10-12 November 2011.
- Cvetana Krstev, Gordana Pavlović-Lažetić, Duško Vitas and Ivan Obradović. 2004. Using textual resources in developing Serbian wordnet. *Romanian Journal of Information Science and Technology*, 7/1-2, pp 147-161.
- Nikola Ljubešić and Darja Fišer, submitted. Extracting translation equivalents for polysemous words from comparable corpora.
- Benoît Sagot and Darja Fišer, 2012. Automatic Extension of WOLF. *Proc. of the 6th International Global Wordnet Conference (GWC'12)*, Matsue, Japan-January, 9-13, 2012.
- Benoît Sagot and Darja Fišer, submitted. Cleaning Noisy Wordnets.
- Špela Vintar. 2011. Samodejno odkrivanje semantičnih premikov v prevodih. *Proc. of 30th simposium Obdobja, Ljubljana, Slovenia*, November 19-19, 2011.

Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base

Aitor González Agirre, Egoitz Laparra, German Rigau

Basque Country University

Donostia, Basque Country

{aitor.gonzalez, egoitz.laparra, german.rigau}@ehu.es

Abstract

This paper describes the upgrading process of the Multilingual Central Repository (MCR). The new MCR uses WordNet 3.0 as Interlingual-Index (ILI). Now, the current version of the MCR integrates in the same EuroWordNet framework wordnets from five different languages: English, Spanish, Catalan, Basque and Galician. In order to provide ontological coherence to all the integrated wordnets, the MCR has also been enriched with a disparate set of ontologies: Base Concepts, Top Ontology, WordNet Domains and Suggested Upper Merged Ontology. We also suggest a novel approach for improving some of the semantic resources integrated in the MCR, including a semi-automatic method to propagate domain information. The whole content of the MCR is freely available.

1 Introduction

Building large and rich knowledge bases is a very costly effort which involves large research groups for long periods of development. For instance, hundreds of person-years have been invested in the development of wordnets for various languages (Fellbaum, 1998; Vossen, 1998; Tufis et al., 2004; K. et al., 2010). In the case of the English WordNet, in more than ten years of manual construction (from 1995 to 2006, that is, from version 1.5 to 3.0), WordNet grew from 103,445 to 235,402 semantic relations¹, which represents a growth of around one thousand new relations per month.

The Multilingual Central Repository (MCR)² (Atserias et al., 2004b) follows the model proposed by the European project EuroWordNet (LE-

2 4003) (Vossen, 1998). The MCR is the result of the European MEANING project (IST-2001-34460) (Rigau et al., 2002), as well as projects KNOW (TIN2006-15049-C03)³ (Agirre et al., 2009), KNOW² (TIN2009-14715-C04)⁴ and several complementary actions associated to the KNOW² project. The original MCR was aligned to the 1.6 version of WordNet. In the framework of the KNOW² project, we decided to upgrade the MCR to be aligned to a most recent version of WordNet.

The previous version of the MCR, aligned to the English 1.6 WordNet version, also integrated the eXtended WordNet project (Mihalcea and Moldovan, 2001), large collections of selectional preferences acquired from SemCor (Agirre and Martínez, 2001) and different sets of named entities (Alfonseca and Manandhar, 2002). It was also enriched with semantic and ontological properties as Top Ontology (Álvez et al., 2008), SUMO (Pease et al., 2002) or WordNet Domains (Magnini and Cavaglià, 2000).

The new MCR integrates wordnets of five different languages, including English, Spanish, Catalan, Basque and Galician. This paper presents the work carried out to upgrade the MCR to new versions of these resources. By using technology to automatically align wordnets (Daudé et al., 2003), we have been able to transport knowledge from different WordNet versions. Thus, we can maintain the compatibility between all the knowledge bases that use a particular version of WordNet as a sense repository. However, most of the ontological knowledge have not been directly ported from the previous version of the MCR.

Furthermore, WordNet Domains⁵ was generated semi-automatically and has never been verified completely. Additionally, it was aligned to

¹Symmetric relations are counted only once.

²<http://adimen.si.ehu.es/web/MCR>

³<http://ixa.si.ehu.es/know>

⁴<http://ixa.si.ehu.es/know2>

⁵<http://wndomains.fbk.eu/>

WordNet 1.6. Thus, one goal of this work is the automatic construction of a new semantic resource derived from WordNet Domains and aligned to WordNet 3.0.

To assist in the correction and maintenance of the integrated resources in the MCR, we also adapted and enhanced the Web EuroWordNet Interface (WEI) in both consult and edit modes.

2 Multilingual Central Repository 3.0

The first version of the MCR was built following the model proposed by the EuroWordNet project. The EuroWordNet architecture includes the InterLingual Index (ILI), a Domain Ontology and a Top Ontology (Vossen, 1998).

Initially most of the knowledge uploaded into the MCR was aligned to WordNet 1.6 and the Spanish, Catalan, Basque and Italian WordNet and the MultiWordNet Domains, were using WordNet 1.6 as ILI (Bentivogli et al., 2002; Magnini and Cavaglià, 2000). Thus, the original MCR used Princeton WordNet 1.6 as ILI. This option also minimized side effects with other European initiatives (Balkanet, EuroTerm, etc.) and wordnet developments around Global WordNet Association. Thus, the Spanish, Catalan and Basque wordnets as well as the EuroWordNet Top Ontology and the associated Base Concepts were transported from its original WordNet 1.5 to WordNet 1.6 (Atserias et al., 1997; Benítez et al., 1998; Atserias et al., 2004a).

The release of new free versions of Spanish and Galician wordnets aligned to Princeton WordNet 3.0 (Fernández-Montraveta et al., 2008; Xavier et al., 2011) brought with it the need to update the MCR and transport all its previous content to a new version using WordNet 3.0 as ILI. Thus, as a first step, we decided to transport Catalan and Basque wordnets and the ontological knowledge: Base Concepts, SUMO, WordNet Domains and Top Ontology.

2.1 Upgrading from 1.6 to 3.0

This section describes the process carried out for adapting the MCR to ILI 3.0. Due to its size and complexity, all this process have been mainly automatic.

To perform the porting between the wordnets 1.6 and 3.0 we have followed a similar process to the one used to port the Spanish and Catalan versions from 1.5 to 1.6 (Atserias et al., 2004a).

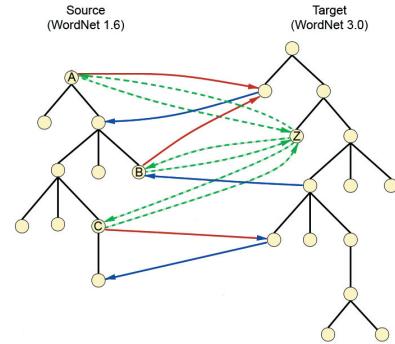


Figure 1: Example of a multiple intersection in the mapping between two versions of WordNet.

Upgrading ILI: The algorithm to align wordnets (Daudé et al., 2000; Daudé et al., 2001; Daudé et al., 2003) produces two mappings for each POS, one in each direction (from 1.6 to 3.0, and from 3.0 to 1.6). To upgrade the ILI, different approaches were applied depending on the POS.

For nouns, those synsets having multiple mappings from 1.6 to 3.0 were checked manually (Pociello et al., 2008).

For verbs, adjectives and adverbs, for those synsets having multiple mappings, we took the intersection between the two mappings (from 1.6 to 3.0, and from 3.0 to 1.6).

Upgrading WordNets: Finally, using the previous mapping, we transported from ILI 1.6 to ILI 3.0 the Basque (Pociello et al., 2008) and Catalan (Benítez et al., 1998) wordnets. The English WordNet was uploaded directly from the source files while the Spanish (Fernández-Montraveta et al., 2008) and Galician (Xavier et al., 2011) wordnets were directly uploaded from their database dumps.

It is possible to have multiple intersections for a source synset. When multiple intersections collapsed into the same target synset, we decided to join the set of variants from the source synsets to the target synset.

Figure 1 shows an example of this particular case (the intersections are displayed as dot lines). Therefore, the variants of the synsets A, B and C of WordNet 1.6 will be placed together in the synset Z of WordNet 3.0.

Upgrading Base Concepts: We used *Base Concepts* directly generated for WN 3.0⁶

⁶<http://adimen.si.ehu.es/web/BLN>

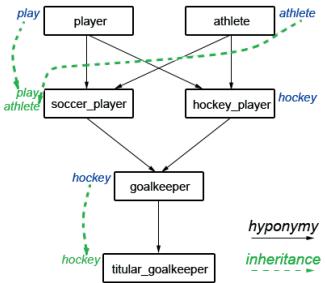


Figure 2: Example of a multiple intersection in the mapping between two versions of WordNet.

(Izquierdo et al., 2007).

Upgrading SUMO: SUMO has been directly ported from version 1.6 using the mapping. Those unlabelled synsets have been filled through inheritance. The ontology of the previous version is a modified version of SUMO, trimmed and polished, to allow the use of first-order theorem provers (like *E-prover* or *Vampire*) for formal reasoning, called AdimenSUMO⁷. The next step is to update AdimenSUMO using the latest version of SUMO for WordNet 3.0 (available on the website of SUMO)⁸.

Upgrading WordNet Domains: As SUMO, what is currently in the MCR has been transported directly from version 1.6 using the mapping. Again, those unlabelled synsets have been filled through inheritance. In addition, new versions have been generated using graph techniques (see Section 4 for a detailed description of the process).

Upgrading the Top Ontology: Similar to SUMO and WordNet Domains, what is currently available in the MCR has been transported directly from version 1.6 using the mapping. Once more, those unlabelled synsets have been filled through inheritance. It remains to check the incompatibilities between labels following (Álvez et al., 2008).

An example of how to perform the process of inheritance used for SUMO, WordNet Domains and Top Ontology is shown in Figure 2. The example is presented for domains, but it can be applied to the other two cases.

Figure 2 shows a sample hierarchy where each node represents a synset. The domains are displayed on the sides. The inherited domain labels

are highlighted using dot lines. In this specific example synset *soccer_player* inherits labels *play* and *athlete* from its hypernyms *player* and *athlete*, respectively. Note that synset *hockey_player* does not inherit any label form its hypernyms because of it owns a domain (*hockey*). Similarly, synset *goalkeeper* does not inherit domains coming from the synset *soccer_player*. Finally, synset *titular_goalkeeper* inherits *hockey* domain (but neither *play* nor *athlete* domains).

Thus, some of the current content of the MCR will require a future revision. Fortunately, by cross-checking its ontological knowledge most of these errors can be easily detected.

2.2 Web EuroWordNet Interface

WEI is a web application that allows consulting and editing the data contained in the MCR and navigating through them. Consulting refers to exploring the content of the MCR by accessing words, a synsets, a variants or ILIs. The interface presents different searching parameters and displays the query results. The different searching parameters are:

- **Item:** a value to search for, it can be a Word, a Synset a Variant or an ILI.
- **Item type:** the type of item to search for: Word, a Synset a Variant or an ILI.
- **PoS:** the item's grammatical category or Part of Speech: Nouns, Verbs, Adjectives, Adverbs.
- **Search:** the type of search and subsearches (which are dynamically loaded from the database): Synonyms, Hyponyms, etc.
- **WordNet Source:** the WordNet from which navigate.
- **Navigation WordNet:** the WordNet to which navigate.
- **Gloss:** if selected it shows the glosses of the Synsets.
- **Score:** if selected, it shows the confidence factor.
- **Rels:** if selected, it shows information about the relations that each Synset has in all the target languages.

⁷<http://adimen.si.ehu.es/web/adimenSUMO>

⁸<http://www.ontologyportal.org/>

- **Full:** if selected, makes a recursive search.
- **Target WordNets:** the target WordNets of our search.

2.3 Automatic translations

The new version of WEI is able to use *Automatic Translation Web Services* for translating automatically the glosses and examples from other wordnets. This new feature helps users to complete and/or improve the gloss or examples of a given WordNet more quickly. Both glosses and examples are taken from the original English WordNet and translated to the target language. Suggestions for glosses and/or examples will appear below the existing ones, and may choose the most appropriate. In the current version, the translations of the glosses and examples are translated only from English (despite the possibility of translating from any available source).

2.4 Marks for synsets and variants

In the new version of WEI it is possible to assign a mark to a variant or synset to indicate special properties. We can also write a small note or comment to explain better the reason to assign that mark.

The available marks are the following:

- Variant marks:
 - DUBLEX: For those variant with dubious lexicalization.
 - INFL: Indicates that the variant is a inflected one.
 - RARE: Old fashioned or rarely used variant.
 - SUBCAT: Subcategorization.
 - VULG: For those variants that are vulgar, rude, or offensive.
- Synset marks:
 - GENLEX: Non-lexicalized general concepts that are introduced to better organize the hierarchy.
 - HYPLEX: Indicates that the hypernym has identical lexicalization.
 - SPECLEX: Domain specific terms that should be checked.

2.5 User management

We also included a new user access control to WEI. The previous user access control to WEI was

carried out using Apache, in the Operating System. This implies that the access control and user management was done outside WEI. The MCR is being edited in a distributed way. Several research groups are editing the MCR in some of the languages. Each group has different users. Thus, the responsibility of managing the users is also distributed.

3 Current state of the MCR

In this section provide some information about the current state of the MCR, including the progress over the English WordNet.

Tables 1, 2 and 3 present respectively the current number of synsets and variants, the number of glosses and the number of examples of each wordnet per PoS.

4 A proposal for upgrading WordNet Domains

WordNet Domains⁹ (WND) is a lexical resource developed at ITC-IRST where synsets have been semi-automatically annotated with one or more domain labels from a set of 165 hierarchically organized domains (Magnini and Cavaglià, 2000). WND allow to reduce the polysemy degree of the words, grouping those senses that belong to the same domain (Magnini et al., 2002).

But the semi-automatic method used to develop this resource was not free of errors and inconsistencies. By cross-checking the ontological content of the MCR it is possible to find some of these problems. For instance, noun synset <diver_1 frogman_1 underwater_diver_1> defined as *someone who works underwater* has domain *history* because it inherits from its hypernym <explorer_1 adventurer_2>.

4.0.1 Domain inheritance

WND was developed using WordNet 1.6. One consequence of the automatic mapping that we used to upgrade version 1.6 to 3.0 is that many synsets were left unlabeled (because there are new synsets, changes in the structure, etc.).

One of the first tasks undertaken has been to fill these gaps. For them, we has carried out a propagation of the labels by inheritance for nominal and verbal synsets. The inherent structure of WordNet for adjectives and adverbs makes this spread

⁹<http://wndomains.fbk.eu/>

WordNet	Nouns	Verbs	Adjectives	Adverbs	Synsets	WN %
EngWN3.0	147,360	25,051	30,004	5,580	118,431	100%
SpaWN3.0	40,009	11,107	7,005	1,106	59,227	50%
CatWN3.0	51,598	11,577	7,679	2	46,027	39%
EusWN3.0	41,071	9,472	148	0	30,615	26%
GalWN3.0	9,114	1,413	4,866	0	9,320	8%

Table 1: Current number of synsets and variants of each WN.

WordNet	Nouns	Verbs	Adjectives	Adverbs	Synsets	WN %
EngWN3.0	82,379	13,767	18,156	3,621	117,923	100%
SpaWN3.0	13,014	3,469	1,965	687	19,135	16%
CatWN3.0	6,289	44	840	1	7,174	6%
EusWN3.0	2,854	78	0	0	2,932	2%
GalWN3.0	4,997	2	3,111	0	8,111	7%

Table 2: Current number of glosses of each WN.

not trivial. Therefore, this simple process has been carried out only for nouns and verbs.

We have worked exclusively on those synsets that had no labels at all. We inherited the labels from its hypernyms. If a synset has more than one hypernym, the domain labels are taken from all of them. We used a small list of incompatible labels to detect incompatibilities. Therefore, the same synset can not be both *factotum* and *biology*, or *animals* and *plants*.

This process increased our domain information by nearly a 18-19%, as shown in Tables 4 and 5:

PoS	Before	After	Increase
Nouns	66,595	83,286	+25%
Verbs	12,219	14,224	+16%
All	100,315	119,011	+19%

Table 4: Number of synsets with domain labels.

PoS	Before	After	Increase
Nouns	87,938	108,665	+24%
Verbs	13,026	15,051	+16%
All	124,551	146,899	+18%

Table 5: Total number of domain labels.

However, this process may also have propagated inappropriate domain labels to unlabeled synsets. It remains for future research an accurate evaluation of this new resource.

In the next section we present some examples using a new graph-based method for propagating

domain labels through WordNet. Additionally, the method can also be used to detect anomalies in the original WND labels.

4.0.2 A new graph based method

UKB¹⁰ algorithm (Agirre and Soroa, 2009) applies personalized PageRank on a graph derived from a wordnet. This algorithm has proven to be very competitive on Word Sense Disambiguation tasks and it is easily portable to other languages that have a wordnet (Agirre et al., 2010). Now, we present a novel use of the UKB algorithm for propagating information through a wordnet structure.

Given an input context, '*ukb-ppv*' (*Personalized PageRank Vector*) algorithm outputs a ranking vector over the nodes of a graph, after applying a *Personalized PageRank* over it. We just need to use a wordnet as a knowledge base and pass to the application the contexts we want to process, performing a kind of *spreading activation* through the WordNet structure.

As context we use those synsets labelled with a particular domain. Thus, for each of the 169¹¹ domain labels included in the MCR we generate a context. Each file contains the list of offsets corresponding to those synsets with a particular domain label. After creating the context file, we just need to execute '*ukb-ppv*' that returns a ranking of weights for each WordNet synset with respect to that particular domain.

¹⁰<http://ixa2.si.ehu.es/ukb/>

¹¹Excluding *factotum* labels.

WordNet	Nouns	Verbs	Adjectives	Adverbs	Synsets	WN %
EngWN3.0	10,433	11,583	15,615	3,674	41,305	100%
SpaWN3.0	478	27	195	967	700	2%
CatWN3.0	2,103	46	368	0	2,517	6%
EusWN3.0	2,377	0	0	0	2,377	6%
GalWN3.0	270	2	4,291	0	4,563	11%

Table 3: Current number of examples of each WN.

Once made the process for all domains we have weights for each synset and for each of the domains. Therefore, we know which are the highest weights for each domain and the highest weights for each synset. This allows us to estimate which synsets are more representative of each domain and which domains are best for each synset.

Basically, what we do is to mark some synsets with a domain (using the labels we already know from the original porting process) and use the wordnet graph to propagate the new labelling. We work on the assumption that a synset directly related to several synsets labelled with a particular domain (i.e. *biology*) would itself possibly be also related somehow to that domain (i.e. *biology*). Therefore, it makes no sense to use the domain *factotum* for this technique.

Table 6 shows the first ten domains and weights resulting from the application of this method on synset <diver_1 frogman_1 underwater_diver_1>. The suggestions of the algorithm seems to improve the current labeling because it suggests *sub* (possibly the best one) and *diving* (possibly, the second best option). Moreover, the method suggests the wrong label with a much lower weight.

Weight	Domain
0.0144335:	sub
0.0015939:	diving
0.0001725:	swimming
0.0001297:	history
0.0000557:	nautical
0.0000529:	fashion
0.0000412:	jewellery
0.0000315:	ethnology
0.0000274:	archaeology
0.0000204:	gas

Table 6: PPV weight rankings for sense *diver*¹_{*n*}.

Table 7 shows the first ten domains and weights resulting from the application of this method on

synset <pornography_1 porno_1 porn_1 erotica_1 smut_5> defined as *creative activity (writing or pictures or films etc.) of no literary or artistic value other than to stimulate sexual desire* and labelled with the domain *law*. The suggestions of the algorithm seems to improve the current labeling because it suggests *sexuality* (possibly the best one) and *cinema* (possibly, the second best option). Moreover, the wrong label disappears.

Method 3	
Weight	Domain
0.000123453:	sexuality
0.000112444:	cinema
0.000077780:	theatre
0.000075525:	painting
0.000062377:	telecommunication
0.000060640:	publishing
0.000050370:	psychological_features
0.000047003:	photography
0.000046853:	artisanship
0.000040458:	graphic_arts

Table 7: PPV weight rankings for sense *porno*¹_{*n*}.

5 Concluding Remarks and Future Directions

As a result of this work, the current version of the MCR consistently maintains new wordnet versions for five languages (English, Spanish, Catalan, Basque and Galician), and the ontological knowledge from WordNet Domains, Top Ontology and SUMO.

In particular, the main contributions of our work can be summarized as follows:

We have created a new version of the MCR using WordNet 3.0 as ILI.

We have improved the existing Web EuroWordNet Interface (WEI) (both consult and edit interfaces) to work with the new version of the MCR. Now, the interface includes automatic translation

facilities, making it easier and faster the development of the resources integrated into the MCR. We also added new editing facilities for recording new linguistic information associated to the variants and synsets.

We have uploaded into the new version of the MCR the English WordNet 3.0, the new Spanish WordNet 3.0 (Fernández-Montraveta et al., 2008) and a new Galician WordNet 3.0.

We have used a complete mapping from WordNet 1.6 to WordNet 3.0 (covering not only nouns, but verbs, adjectives and adverbs) to transport the Basque and Catalan wordnets and the ontological knowledge from the existing version of the MCR (using WordNet 1.6 as ILI) to the new MCR version (using WordNet 3.0 as ILI).

We have applied a very simple strategy to complete the ontological information by exploiting basic inheritance mechanisms. This process has been applied to WordNet Domains, Top Ontology and SUMO.

We have also investigated a new approach for consistently propagating domain formation through the WordNet structure by exploiting a well-known graph algorithm using UKB. Although an exhaustive empirical evaluation should be addressed in a near future, a preliminary review of the new resources created using this process presents very interesting insights for future research.

Obviously, further investigation is needed to assess the quality of the new labelling of WordNet Domains. We plan to evaluate the quality of these new resources indirectly by comparing their performance on a common Word Sense Disambiguation task. We would also like to continue studying different ways for selecting the most appropriate set of domain labels per synset. We also plan to derive domain information from Wikipedia by exploiting WordNet++ (Navigli and Ponzetto, 2010).

The whole content of the MCR and the new WEI is freely available¹².

Moreover, the maintenance of this type of resources is continuous, and all the integrated knowledge should be constantly updated and revised.

Acknowledgments

We thank the IXA NLP group from the Basque Country University. This work was been pos-

sible thanks to the support of the FP7 PATHS project (ICT-2010-270082) and the Spanish national projects KNOW (TIN2006-15049-C03) and KNOW2 (TIN2009-14715-C04-04).

References

- Agirre, E., Cuadros, M., Rigau, G., and Soroa, A. (2010). Exploring Knowledge Bases for Similarity. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC10)*. European Language Resources Association (ELRA). ISBN: 2-9517408-6-7. Pages 373–377.”.
- Agirre, E. and Martinez, D. (2001). Knowledge sources for Word Sense Disambiguation. In *Proceedings of the Fourth International Conference TSD 2001, Plzen (Pilsen), Czech Republic*. Published in the Springer Verlag Lecture Notes in Computer Science series. Václav Matousek, Pavel Mautner, Roman Mouncek, Karel Tauser (eds.) Copyright Springer-Verlag. ISBN: 3-540-42557-8. ”.
- Agirre, E., Rigau, G., Castellón, I., Alonso, L., Padró, L., Cuadros, M., Climent, S., and Coll-Florit, M. (2009). KNOW: Developing large-scale multilingual technologies for language understanding. *Procesamiento del Lenguaje Natural*, (43):377–378.
- Agirre, E. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.
- Alfonseca, E. and Manandhar, S. (2002). Distinguishing concepts and instances in WordNet. In *Proceedings of the first International Conference of Global WordNet Association*, Mysore, India.
- Álvez, J., Atserias, J., Carrera, J., Climent, S., Laparra, E., Oliver, A., and Rigau, G. (2008). Complete and Consistent Annotation of WordNet using the Top Concept Ontology. In *Proceedings of the the 6th Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech (Morocco).
- Atserias, J., Climent, S., Farreres, J., Rigau, G., and Rodríguez, H. (1997). Combining Multiple Methods for the Automatic Construction of Multilingual WordNets. In *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP'97)*, Tzigov Chark, Bulgaria.
- Atserias, J., Rigau, G., and Villarejo, L. (2004a). Spanish WordNet 1.6: Porting the Spanish Wordnet across Princeton versions. In *LREC'04*.
- Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Vossen, P., and Magnini, B. (2004b). The MEANING Multilingual Central Repository. In *Proceedings of the Second International Global WordNet Conference (GWC'04)*.

¹²<http://adimen.si.ehu.es/web/MCR>

- Bentivogli, L., Pianta, E., and Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India.
- Benítez, L., Cervell, S., Escudero, G., López, M., Rigau, G., and Taulé, M. (1998). Methods and Tools for Building the Catalan WordNet. In *Proceedings of ELRA Workshop on Language Resources for European Minority Languages*, Granada, Spain.
- Daudé, J., Padró, L., and Rigau, G. (2000). Mapping WordNets Using Structural Information. In *38th Annual Meeting of the Association for Computational Linguistics (ACL'2000)*, Hong Kong.
- Daudé, J., Padró, L., and Rigau, G. (2001). A Complete WN1.5 to WN1.6 Mapping. In *Proceedings of NAACL Workshop 'WordNet and Other Lexical Resources: Applications, Extensions and Customizations' (NAACL'2001)*, Pittsburg, PA, USA.
- Daudé, J., Padró, L., and Rigau, G. (2003). Making Wordnet Mappings Robust. In *Proceedings of the 19th Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN*, Universidad Universidad de Alcalá de Henares, Madrid, Spain.
- Fellbaum, C. (1998). *WordNet. An Electronic Lexical Database*. Language, Speech, and Communication. The MIT Press.
- Fernández-Montraveta, A., Vázquez, G., and Fellbaum, C. (2008). *Text Resources and Lexical Knowledge*, volume 33 of *Text, Translation, Computational Processing*, chapter The Spanish Version of WordNet 3.0, pages 175–182. Mouton de Gruyter.
- Izquierdo, R., Suárez, A., and Rigau, G. (2007). Exploring the Automatic Selection of Basic Level Concepts. In *Proceedings of RANLP*.
- K., R., S., T., T., C., V., S., and H., I. (2010). WNMS: Connecting the Distributed WordNet in the Case of Asian WordNet. In *Proceedings of the 5th Global WordNet Conference (GWC'10)*, Mumbai (India).
- Magnini, B. and Cavaglià, G. (2000). Integrating subject field codes into WordNet. In *Proceedings of the Second Internatigional Conference on Language Resources and Evaluation (LREC)*, Athens, Greece.
- Magnini, B., Satrapparava, C., Pezzulo, G., and Gliozzo, A. (2002). The Role of Domains Informations. In *In Word Sense Disambiguation*, Treto, Cambridge.
- Mihalcea, R. and Moldovan, D. (2001). eXtended WordNet: Progress Report. In *Proceedings of NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pages 95–100, Pittsburg, PA, USA.
- Navigli, R. and Ponzetto, S. P. (2010). Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Up psala, Sweden.
- Pease, A., Niles, I., and Li, J. (2002). The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In *AAAI-2002*.
- Pociello, E., Gurrutxaga, A., Agirre, E., Aldezarbal, I., and Rigau, G. (2008). WNTERM: Combining the Basque WordNet and a Terminological Dictionary. In *6th international conference on Language Resources and Evaluation, LREC'08*.
- Rigau, G., Magnini, B., Agirre, E., Vossen, P., and Carroll, J. (2002). MEANING: A Roadmap to Knowledge Technologies. In *Proceedings of COLING Workshop A Roadmap for Computational Linguistics*, Taipei, Taiwan.
- Tufis, D., Cristea, D., and Stamou, S. (2004). Balka Net: Aims, Methods, Results and Perspectives. A General Overview. *Romanian Journal on Science Technology of Information. Special Issue on Balkan net*, 7(3–4):9–44.
- Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.
- Xavier, G. G., Clemente, X. M. G., Pereira, A. G., and Lorenzo, V. T. (2011). Galnet: WordNet 3.0 do galego. *Linguamática*, 3(1):61–67.

New WordNet-based semantic relatedness measurement

Using new information content metric and k-means clustering algorithm

Mohamed Ali Hadj Taieb

MIRACL Laboratory
FSEGS, B.P. 1088,
3018 Sfax, Tunisia
mohamedali.hadjtaieb@gmail.com

Mohamed Tmar

MIRACL Laboratory
ISIMS, Ons City,
3021 Sfax, Tunisia
mohamed.tmar@isimsf.rnu.tn

Mohamed Ben Aouicha

MIRACL Laboratory
ISECS, B.P. 868,
3000 Sfax, Tunisia
mohamed.benaouicha@irit.fr

Abdelmajid Ben Hamadou

MIRACL Laboratory
ISIMS, Ons City,
3021 Sfax, Tunisia
abdelmajid.benhamadou@isimsf.rnu.tn

Abstract

Semantics constitute one of the major stakes in the information retrieval (IR) system evolution. Taking semantics into account passes by the use of external semantic resources coupled with the initial documentation on which it is necessary to have semantic similarity measurements to carry out comparisons between concepts. This paper presents a new approach for measuring semantic relatedness between words and concepts. It combines a new information content (IC) metric using the WordNet thesaurus and the k-means clustering algorithm. Specifically, the proposed method offers a thorough use of the relation hypernym/hyponym ("is a" taxonomy) without external corpus statistical information. Mainly, we use the subgraph formed by hypernyms of the concerned concept which inherits the whole features of its hypernyms. Moreover, we exploit the significant depth in "is a" hierarchical structure. The 3-means clustering algorithm is used to exceed the problem of very fine granularity. When tested on a common data set of word pair similarity ratings, the proposed approach outperforms other computational models. It gives the highest correlation value 0.72 with a benchmark based on human similarity judgments and especially a large dataset composed of 200 Finkelstein word pairs.

1 Introduction

Semantic similarity is an important topic in Natural Language Processing (NLP) and IR. It is commonly accepted that the use of semantic resources like ontologies and taxonomies of con-

cepts improves the performances of the IR systems (Renard et al., 2010). Closely related works to this study are those that were aligned with the thread of our discussion. Indeed, application areas of semantic similarity include word sense disambiguation (WSD) (Resnik, 1999), detection and correction of word spelling errors (malapropisms) (Budanitsky and Hirst, 2001) and semantic indexing (Zargayouna, 2004).

To quantify the concept of similarity between words, some ideas have been put forth by researchers, most of which rely heavily on the knowledge available in lexical knowledge bases like WordNet¹.

There are mainly two approaches to compute semantic similarity. The first approach is making use of a large corpus or word definitions and gathering statistical data from these sources to estimate a score of semantic similarity, which we call text-based approach. The second approach makes use of the relations and the hierarchy of a thesaurus, such as Word-Net, which we call structure-based approach.

In text-based approach, word relationships are often derived from their co-occurrence distribution in a corpus (Grefenstette, 1992). Gloss overlap, introduced by Lesk (1986) and extended gloss overlap, introduced by Banerjee and Pedersen, are other instances of this approach. The latter is a measure that determines the relatedness of concepts proportional to the extent of overlap of their WordNet glosses (Banerjee and Pedersen, 2003). Besides gloss vector measure of semantic relatedness, introduced by Pedersen and

¹ <http://wordnet.princeton.edu>

Patwardhan, is based on second order co-occurrence vectors in combination with the structure and content of WordNet, a semantic network of concepts (Patwardhan and Pedersen, 2006).

In structure-based approach, first studies date back to Quilian's semantic memory model (1968) and MacQueen studies (1967), where the number of hops between nodes of concepts in the hierarchical network specifies the similarity or difference of concepts. Wu and Palmer's semantic similarity measure (1994) was based on the path length between concepts located in taxonomy. Also, the similarity measure of Leacock and Chodorow is based on the shortest path length between two concepts in is-a hierarchy (Leacock and Chodorow, 1998).

In combining two approaches, Resnik (1995) introduced a new factor of relatedness called information content (IC). The Similarity measures of Resnik, Jiang and Conrath (1997) and Lin (1998) all rely on the IC values assigned to the concepts in an is-a hierarchy, but their usage of IC has few differences.

Using a different approach Hirst G. and St-Onge assign relatedness scores to words rather than word senses. They set different weights for different kinds of links in a semantic network, and uses those weights for edge counting (Hirst and St-Onge, 1998).

In this paper, we first introduce a new method for computing IC of concepts in a hierarchical structure. We will show that this method uses only the hierarchical structure and not corpus to determine IC. Furthermore, information content obtained from this method includes the subgraph formed by hypernyms of the concerned concept and the depth from root to target concept. Then we use formula that is similar to Lin formula for measuring similarity coupled with k-means clustering algorithm. Finally our similarity measure is evaluated against a benchmark set of human similarity ratings, and demonstrates that the proposed measure significantly outperformed traditional similarity measures.

In section 2 we describe WordNet, which was used in our method. Section 3 describes the approaches of information content metric using WordNet taxonomic structure as semantic resource. Section 4 presents our novel information content metric using only the WordNet taxonomic structure. Section 5 defines the novel semantic similarity measurement using k-means algorithm. Section 6 presents an evaluation of the similarity measure against human similarity judgments benchmarks. Finally, the paper con-

cludes in Section 7 that, based on the benchmark data set, our measure outperforms existing measures, and we expose future directions of this study.

2 WordNet

The WordNet (Fellbaum, 1998) is an electronic lexical database created at Princeton University in 1990. The WordNet organizes the lexical information in meanings (senses) and synsets (set of words – sentence(s) - describing the meaning of the word in a specific context). Each synset has a gloss that defines the concept of the word. For example the words *car*, *auto*, *automobile*, and *motorcar* is a synset that represents the concept defined by gloss: *four wheel Motor vehicle, usually propelled by an internal combustion engine*. Many glosses have examples of usages associated with them, such as "*he needs a car to get to work*." What makes WordNet remarkable is the existence of various relations between the word forms. These semantic relations for nouns include: Hyponym/Hypernym (IS-A/ HAS A), Meronym/Holonym (Part-of / Has-Part), Meronym/Holonym (Member-of / Has-Member) and Meronym/Holonym (Substance-of / Has-Substance).

The hyponymy relation, used in our method, is a relation that organizes nouns and verbs into a lexical inheritance system. It is an *is-a* hierarchy. In this hierarchical system, a subordinate term inherits the basic features from the superordinate term and adds its own distinguishing features to form its meaning. Hence, the organization of the WordNet allows us to start from a topmost node or from an intermediate one and to go up or down finding the broader or narrower (more specific) meanings and then use them in a variety of ways.

3 The WordNet-based Information Content Metric

The IC concept was introduced for the first time by Resnik (Resnik, 1995). Following the standard argumentation of information theory (Ross, 1976), the information content of a concept c can be quantified as negative the log likelihood, $-\log p(c)$. Notice that quantifying information content in this way makes intuitive sense in this setting: as probability increases, informativeness decreases, so the more abstract a concept, the lower its information content.

3.1 Previous Information Content Based Approach: Statistically Analyzing Corpora

The IC is an important dimension of word knowledge when assessing the similarity of two terms or word senses. The conventional way of measuring the IC of word senses is to combine knowledge of their hierarchical structure from ontology like WordNet with statistics on their actual usage in text as derived from a large corpus.

Many researchers consider statistical figures to compute IC value. They assign a probability to a concept in taxonomy based on the occurrence of target concept in a given corpus. For Resnik (1995), frequencies of concepts in taxonomy were estimated using noun frequencies from the Brown Corpus of American English (Francis and Kucera, 1982), a large (1,000,000 word) collection of text across genres ranging from new articles to science fiction. Each noun that occurred in the corpus was counted as an occurrence of each taxonomic class containing it. The IC value is then calculated by negative log likelihood formula as follow:

$$IC(c) = -\log(p(c)) \quad (1)$$

Where c is a concept and p is the probability of encountering c in a given corpus. If sense-tagged text is available, frequency counts of concepts can be attained directly, since each concept will be associated with a unique sense. If sense tagged text is not available, it will be necessary to adopt an alternative counting scheme. Resnik (1995) suggests counting the number of occurrences of a word type in a corpus, and then dividing that count by the number of different concepts/senses associated with that word. This value is then assigned to each concept. For example, suppose that the word type *bank* occurs 20 times in a corpus, and that there are two concepts associated with this type in the hierarchy, one for *river bank* and the other for *financial bank*. Each of these concepts would receive a count of 10. If the occurrences of bank were sense tagged then the relevant counts could simply be assigned to the appropriate concept. Resnik showed that semantic similarity depends on the amount of information that two concepts have in common, this shared information is given by the *Most Specific Common Abstraction* (MSCA) that subsumes both concepts.

Therefore we must first discover the MSCA and then shared information is equal to the IC value

of the MSCA. If MSCA does not exist then the two concepts are maximally dissimilar.

3.2 Previous Information Content Based Approach: WordNet Taxonomic Structure

As was made clear in the previous paragraph, IC is obtained through statistical analysis of corpora, from where probabilities of concepts occurring are inferred. Other authors feel that WordNet can also be used as a statistical resource with no need for external ones. Moreover, they argue that the WordNet taxonomy may be innovatively exploited to produce the IC values needed for semantic similarity calculations. Their methods of obtaining IC values rest on the assumption that the taxonomic structure of WordNet is organized in a meaningful and principled way.

Nuno and al. (2004) present a completely intrinsic measurement of IC which is connected only to the hierarchical structure of WordNet. This minimizes the complexity of calculation while getting rid of the time necessary to the treatment of a corpus in order to extract the probabilities from the concepts. Their idea coincides better with the human judgments than by using extrinsic measurements of IC which employ an external corpus. The information content of a concept c depends amongst concept which it subsumes. The computation formula is the following:

$$IC(c) = 1 - \frac{\log(hypo(c)+1)}{\log(max_{wn})} \quad (2)$$

$hypo(c)$ is a function which returns the hyponyms number of a given concept and max_{wn} is a constant which represents the maximum number of concepts in the WordNet thesaurus (in WordNet 2.0 $max_{wn}=79689$).

Sebti and al. (2008) present a new approach for measuring semantic similarity between words via concepts. Their proposed measure is a hybrid system based on using a new Information content metric with the WordNet hierarchical structure. This method implicitly includes the concept of depth of a target concept. Figure 1 explains the process followed in order to calculate the IC of a given concept. As we notice, this method is based on the number of direct hyponyms of each concept pertaining to the initial way of the root until the target concept.

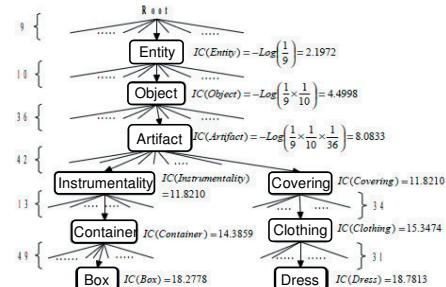


Figure 1. Example of IC computing for some concepts using Sebti method.

For a better understanding of this method, we present the following equation that represents the IC calculated for the concept *Box* in relation to figure 1:

$$CI(Box) = -\log \left(\frac{1}{9} \times \frac{1}{10} \times \frac{1}{36} \times \frac{1}{42} \times \frac{1}{13} \times \frac{1}{49} \right) = 18.2778$$

4 Our IC Metric

Hence, we present a novel IC metric that is completely derived from WordNet without the need for external resources from which statistical data is gathered. In (Hadj Taieb *et al.*, 2011), we have discussed the limits of approaches which are already cited in the previous section (Nuno and Sebti).

The two principal ideas in the creation of our new method are explained in the following paragraphs.

4.1 The Taxonomic Relation «is a» or Hypernym/Hyponym

In this hierarchical system, a subordinate concept inherits the basic features from the superordinate concept and adds its own specific features to form its meaning. Thus, a concept is an accumulation of the propagated information of an ancestor to another by adding specificity to each descendant. Therefore, a concept depends strongly on its direct parents and their ancestors. Direct and indirect hypernym relations of a concept *c* from a subgraph which will take part in its IC quantification.

The notation *w#i#j* represents the synset number *i* amongst *j* synsets of the word *w*.

In figure 2 we present an example of two hypernyms subgraphs for concepts *atropine#1#1* (*subgraph(atropine#1#1)=*{a,b,c,d,e,f,g,h,i,j,k,l,m,n,p}) and *obidoxime_chloride#1#1* (*subgraph(obidoxime_chloride#1#1)=*{a,b,d,h,j,l,n,o}). Concerning this example, it is clear the remarkable difference between volumes of both subgraphs. Therefore, the IC of the concept *atropine#1#1* must be largely higher than the IC of the concept *obidoxime chloride#1#1*.

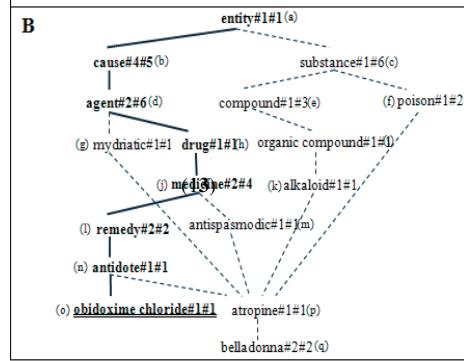
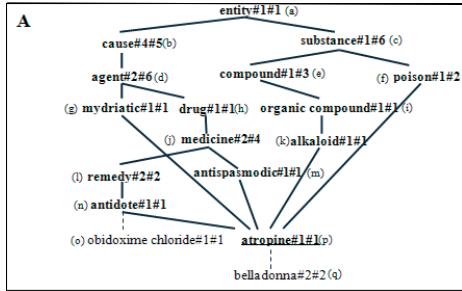


Figure 2. Wordnet fragment that represents subgraphs defining $IC(atropine\#1\#1)$ (Part A) and $IC(obidoxime\#1\#1)$ (Part B). Concepts are labeled by alphabetic letters. Besides, the solid lines in Parts A and B represent respectively the subgraph of *atropine#1#1* and *obidoxime chloride#1#1*.

4.2 The Significant Depth in Taxonomic Structure “is a”

Each concept except the root has one or more hypernyms (up to 6 in WordNet 2.0). These hypernyms do not have the same depth in the taxonomy “is a”. The depth is significant because raising from a level to another generates the data propagation towards the descendants with addition of certain specificities. The transition from concept *c*₁ towards concept *c*₂ using the hyponymy relation does not mean the passage from depth *i* to depth *i+1*. Thus, two concepts, directly connected, do not have necessarily successive depths in “is a” taxonomy (For example, the synsets *poison#1#2* and *atropine#1#1* in the figure 2 are directly connected but the depth is equal to 3 for the first synset and for the second synset, depth=8).

4.3 New IC Method

In this section, we explain our method which quantify hypernyms subgraph of a given concept and take into account the depth influence of each concept and the mean average depth of the subgraph. Therefore, we use these notations:

Hyper(*c*): the set of direct hypernyms of the concept *c*. For example, $Hyper(atropine\#1\#1)=\{n,g,m,k,f\}$ and $Hyper(obidoxime\#1\#1)=\{n\}$.

Hypo(*c*): the number of concepts subsumed by the concept *c*. For example, $Hypo(entity\#1\#1)=74104$ (in WordNet 2.0).

SubGraph(*c*): set of all concepts belonging to hypernyms subgraph modelling the information content of the concept *c*. For example $sub-$

$\text{graph}(atropine\#1\#1)=\{a,b,c,d,e,f,g,h,i,j,k,l,m,n,p\}$.

Depth(c): the maximal depth of the concept c in the WordNet thesaurus. For example, $\text{Depth}(atropine\#1\#1)=8$.

We suppose that we want to compute $IC(Con)$ where Con is a concept.

For each $c \in SubGraph(Con)$, we calculate a score:

$$Score(c) = \left(\sum_{c' \in Hyper(c)} \frac{Depth(c')}{Hypo(c')} \right) \times Hypo(c) \quad (3)$$

Indeed, for each concept $c' \in Hyper(c)$ we calculate a term as follows: $\frac{Hypo(c)}{Hypo(c')} \times Depth(c')$ (term). $Score(c)$ represents the contribution of each concept belonging to $Subgraph(Con)$ on $IC(Con)$.

For noun POS, WordNet 2.0 contains 9 roots (entity#1#1, event#1#4, group#1#3, phenomenon#1#2, possession#2#7, psychological feature#1#1, state#4#8, abstraction#6#6 and act#2#5). Then, if the concept c is one of these 9 concepts, we suppose that we have a principle root that subsumes them to calculate $Score(c)$.

For example, $Score(atropine\#1\#1)$ is equal to the sum of contribution rate of each hypernym ({n,g,m,k,f}) in its score. This contribution rate was calculated thanks to the already cited term. Concerning $Score(atropine\#1\#1)$, the term increases when the hyponyms of $atropine\#1\#1$ represent most of the hyponyms of its hypernyms (hyponyms of *antidote*#1#1, hyponyms of *mydriatic*#1#1 ...).

Moreover, the contribution rate of a hypernym c belonging to $Subgraph(Con)$ depends on its depth. Because, while going down from the root towards any target concept, there is a data enrichment. Therefore, the term contains a multiplication by $Depth(c')$. For example, the concepts *poison*#1#2 ($depth=3$) and *antidote*#1#1 ($depth=7$) are hypernyms of *atropine*#1#1 concept. Inspite, the contribution of *poison*#1#2 concept in $score(atropine\#1\#1)$ is less important.

Finally, we calculate the total IC of the concept Con as follows:

$$IC(Con) = \sum_{c \in SubGraph(Con)} Score(c) \quad (4)$$

5 New Semantic Similarity Measure

In order to evaluate our new approach, we use the Lin formula (1998) in order to measure the semantic similarity between words, we use our novel IC method to quantify the similarity between two concepts c_1 and c_2 . We use the MSCA that return the lowest common subsumer of the both concepts:

$$Sim(c_1, c_2) = \frac{2 \times IC(\text{MSCA}(c_1, c_2))}{IC(c_1) + IC(c_2)} \quad (5)$$

This can be viewed as taking the information content of the intersection of the two concepts (multiplied by 2) and dividing it by their sum, which is analogous to the well-known Dice Coefficient.

Then, we use the maximum function to extract the semantic similarity degree between these words:

$$(6)$$

$$Sim(w_1, w_2) = \max_{(c_1, c_2) \in \text{Syn}(w_1) \times \text{Syn}(w_2)} Sim(c_1, c_2)$$

The counterpart of its important cover is that WordNet is very precise in the sense of definitions. There is a too fine granularity of senses. A word in WordNet can have several meanings i.e. a large number of synsets, (such as *book*: 10 synsets and *library*: 5 synsets then we have 50 possible combinations). Then certain particular meanings lead to false measure of semantic similarity. K-means clustering (MacQueen, 1967) is a commonly used method to automatically partitioning a data set into k groups. It proceeds by selecting k initial cluster centers and then iteratively refining them as follows:

1. Each instance is assigned to its closest cluster center.
2. Each cluster center is updated to be the mean of its constituent instances.

The algorithm converges when there is no further change in assignment of instances to clusters. As for us, the distance between 2 values x and y is $|x-y|$, and the distance between a value x and a cluster E is $|x-avg(E)|$.

To calculate the semantic relatedness between two words w_1 and w_2 , we start by calculating the semantic similarity of each concept couple $(c_i, c_j) = \text{Syn}(w_1) \times \text{Syn}(w_2)$ where $\text{Syn}(w)$ represents the synsets of the word w . Then, we select non-zero values. Finally, we apply the k-means algorithm to divide values on 3 sets (E_1 , E_2 and E_3). Setting the clusters number parameter at $k=3$ seems optimal as most of the resultant values outperform others tested experimentally. Finally,

the semantic similarity degree will be the maximum mean. So,

(7)

$$Sim(w_1, w_2) = \max_{E_i \in \{E_1, E_2, E_3\}} avg(E_i)$$

Where $avg(E_i)$ is the center value of E_i cluster.

6 Experiments and Results

We would be reasonable to evaluate the performance of machine measurements of semantic similarity between concepts by comparing them to human ratings on the same setting. The simplest way to implement this is to set up an experiment to rate the similarity of a set of word pairs, and examine the correlation between human judgments and machine calculations. To make our experimental results comparable with other previous experiments, we use some benchmark sets of noun pairs.

Rubenstein and Goodenough (1965) obtained “synonymy judgments” of 51 human subjects on 65 pairs of words. The pairs ranged from “highly synonymous” to “semantically unrelated”. Subjects were asked to rate them on the scale of 0.0 to 4.0 according to their similarity of meaning. Miller and Charles (1991) subsequently extracted 30 pairs from the original 65 and then obtained similarity judgments from 38 subjects. Finkelstein and al. (2002)² (Finkelstein et al., 2002) created a large dataset that included the original MC ‘30’ dataset for their study of human association performance. Subjects were asked to rate them on the scale of 0.0 to 10.0 according to their meaning similarity. We extract 200 noun pairs of Finkelstein dataset F* (see appendix) to test our approach on a large dataset (with an average rate equal to 5.96).

To compare our measurement with those proposed by other authors, we used wordnet 2.0 with the free package created by Siddarth Patwardhan and Ted Pederson³ (2004). This package implements some measures described by Leacock and Chodorow (1998), Wu and Palmer (1994), Jiang and Conrath (1997), Banerjee and Pederson (Extended Gloss Overlap) (2003), Lin (1998), Resnik (1995), Patwardhan and Pederson (Gloss Vectors) (2006), Hirst-St-Onge (1998) and LSA⁴ (Latent Semantic Analysis).

Similarity Measure	R&G	M&C	F*
Lin	0.73	0.80	0.41
Jiang et Conrath	0.68	0.70	0.40
Resnik	0.77	0.77	0.53
Leacock et Chodorow	0.82	0.82	0.57
Hirst-St-Onge	0.70	0.68	0.43
Wu et Palmer	0.76	0.74	0.60

² Data from 13 subjects from the wordsimilarity-353 test collection (Finkelstein et al., 2002). Downloadable from:

<http://alfonseca.org/eng/research/wordsim353.html>.

³ Package downloadable via <http://wn-similarity.sourceforge.net/>.

⁴ LSA plateforme accessible via <http://lsa.colorado.edu/>

Gloss Vectors	0.79	0.88	0.48
Extended Gloss Overlap	0.33	0.34	0.31
LSA	0.65	0.73	0.52
Our (Max)	0.81	0.75	0.63
Our (K-means)	0.82	0.76	0.72

Table 1: The coefficients correlation between human similarity judgments (Miller & Charles, Rubenstein & Goodenough and Finkelstein) and the suggested similarity measures.

In the table 1, we obtained for each implementation similarity measure or semantic relatedness scores for the human rated pairs. We follow Resnik (1995) and other authors in summarizing the comparison results by means of correlation coefficient with the reported human ratings for each computational measure.

For R&G and M&C benchmark set of human similarity, our novel measure shows a significantly better correlation ($r=0.82$ and $r=0.76$). But, for a large dataset (200 noun pairs extracted from Finkelstein word similarity test collection), our measure outperforms other methods. Overall, there is a good performance improvement (10%) over the result when the k-means clustering algorithm is used instead of the max method.

7 Conclusion and Future Work

The result obtained using our IC values in the information theoretic formulas seem to have outperformed their homologies which suggests that the initial assumption concerning the taxonomic structure of WordNet is correct. Indeed, the hypernyms subgraph of a concept expresses the IC as best possible manner. K-means clustering algorithm coupled to the novel IC method leads to a semantic relatedness measurement closer to the human judgments.

This approach uses a novel IC metric which tries to quantify the meaning contribution of each concept belonging to the hypernyms subgraph in a target concept. Moreover, it does not rely on corpora analysis, thus we avoid the sparse data problem which is evident in many corpus based approaches.

Future research regarding the information content metric will make use of taxonomies other than WordNet. This will allow us to conclude if our metric generalizes and can be used with other hierarchical knowledge bases.

In future research, we will emphasize the analysis of the textual WordNet definitions to investigate latent features of concepts. Moreover, we will attempt to evaluate our model in specific applications such as Word Senses Disambiguation (WSD). A powerful semantic relatedness measure influences on Semantic Information Retrieval (SIR) system. In (Harrathi and Calabretto, 2010), the authors use the semantic measures in conceptual indexing for semi-structured documents. It exists some information retrieval system that support retrieval by *Semantic Similarity Retrieval Model (SSRM)* (Varelas et al., 2005), a novel informa-

tion retrieval method which is capable for associating documents containing semantically similar (but not necessarily lexically similar) terms. *SSRM* suggests discovering semantically similar terms in documents and queries using term taxonomies and by associating such terms using semantic similarity methods. Therefore, our method can be tested in a semantic retrieval system supported by *SSRM*.

References

- Banerjee S., and Pedersen T. "Extended gloss overlaps as a measure of semantic relatedness". In the proceedings of the Eighteenth International Joint Conference on Artificial Intelligence. Acapulco, Mexico, 2003.
- Budanitsky A. and Hirst G., "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures". Proc. Workshop WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics , Pittsburgh, june 2001, pp.29-34.
- Leacock C. and Chodorow M., "Combinig local context and WordNet similarity for word sense identification", in C. Fellbaum (Ed.), WordNet: An electronic lexical database, MIT Press, 1998,pp. 265-283, 1998.
- Fellbaum C.: "WordNet: An electronic lexical database". The MIT Press, Cambridge MA 1998.
- Finkelstein L., Gabrilovich E., Matias Y., Rivlin E., Solan Z., Wolfman G. and Ruppin E., "Placing search in context: the concept revisited", ACM Transactions on Information Systems, Vol. 20, no 1. pp 116-131, 2002.
- Francis W. N. and Kucera H., "Frequency analysis of english usage: lexicon and grammar". Houghton Mifflin, 1982.
- Miller A. G. and Charles G. W. "Contextual correlates of semantic similarity". Language and Cognitive Processes, 6 (1): 1-28, 1991.
- Hirst G. and St-Onge D. "Lexical chains as representations of context for the detection and correction of malapropisms". In Fellbaum 1998, pp. 305-332, 1998.
- Grefenstette G., "Use of syntactic context to produce term association lists for text retrieval". Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval , SIGIR'92, 1992.
- Hadj Taieb M., Ben Aouicha M., Ben Hamadou A. "Une nouvelle approche de calcul du contenu informationnel pour mesurer la similitude sémantique utilisant WordNet". Position paper, INFORSID'11, Lille, 2011.
- Rubenstein H. and Goodenough J. B., "Contextual correlates of synonymy". Communications of the ACM, 8 (10): 627-633, 1965.
- Jiang J. and Conrath D. "Semantic similarity based on corpus statistics and lexical taxonomy". In Proceedings of International Conference on Research in Computational Linguistics , Taiwan, 1997.
- Lesk M., "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone". In Proceedings of the SIGDOC Conference, Toronto, 1986.
- Lin D., "An information-theoretic definition of similarity". In Proceedings of the 15th International Conference on Machine Learning, Madison, WI, 1998.
- MacQueen J. B. "Some methods for classification and analysis of multivariate observations". Proceedings of the Fifth Symposium on Math, Statistics, and probability (pp. 281-297). Berkeley, CA: University of California Press, 1967.
- Patwardhan S. and Pedersen T. "Using WordNet-based context vectors to estimate the semantic relatedness of concepts". In Proceedings of Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together , EACL, 2006.
- Pedersen T., and Patwardhan S. and Michelizzi J. "WordNet::Similarity - measuring the relatedness of concepts" In: Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-2004) pp. 1024-1025. San Jose, CA. July, 2004.
- Quillian M. "Semantic memory", in M. Minsky (Ed.), Semantic Information Processing, the MIT press, pages 216-270, 1968.
- Renard A., Calabretto S. and Rumpler B.: "Fuzzy semantic matching in (semi-) structured XML documents: indexation of noisy documents". 6th International Conference on Web Information Systems and Technologies (WEBIST 2010), CORDEIRO, J. ed. Valencia, Spain. pp. 253-260. 2010.
- Resnik P. "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language". J. Artificial Intelligence Research, vol. 11, pp. 95-130, 1999.
- Resnik P. "Using information content to evaluate semantic similarity in a taxonomy". In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 448-453, Montreal, 1995.
- Sebt A.. and Barfrouch A. A.: "A new word sense similarity measure in WordNet", Proceedings of the International Multiconference on Computer Science and Information Technologie. Poland, 2008.

Nuno S., Tony V. and Jer, H. "An intrinsic information content metric for semantic similarity in WordNet". In Proceedings of the 16th European Conference on Artificial Intelligence, Valencia, Spain, 22–27, pages 1089–1090, August 2004

Ross S., "A first course in probability". Macmillan, 1976.

Wu Z. and Palmer M., "Verb semantics and lexical selection", Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics, Las Cruces, New Mexico, pages 133-138, 1994.

Zargayouna H., "Contexte et sémantique pour une indexation de documents semi-structurés". ACM

Appendix: The test dataset F*

The total 200 pairs of nouns

investigation	effort	magician	wizard	love	sex	stock	egg
announcement	production	midday	noon	tiger	cat	dollar	yen
announcement	warning	stock	life	money	cash	dollar	buck
television	radio	bird	cock	century	year	five	month
psychology	psychiatry	bird	crane	delay	racism	report	gain
combination	direction	terror	Arafat	cup	tableware	book	paper
implement	tool	food	rooster	news	report	listing	proximity
computer	keyboard	monk	oracle	territory	surface	coast	hill
development	population	Mexico	Brazil	cup	coffee	network	hardware
manslaughter	murder	cucumber	potato	planet	sun	plane	car
accommodation	arrangement	space	chemistry	furnace	stove	seafood	lobster
equipment	phone	shore	woodland	doctor	professor	bed	closet
forest	graveyard	glass	magician	image	surface	liquid	water
substance	cup	coast	forest	record	number	seafood	food
student	professor	money	deposit	mile	kilometer	Arafat	Jackson
experience	music	wood	forest	jaguar	cat	stock	CD
Wednesday	news	tiger	animal	king	queen	game	round
psychology	cognition	tiger	organism	tiger	tiger	start	year
observation	architecture	fuck	sex	fauna	fauna	rooster	voyage
atmosphere	landscape	tiger	jaguar	sign	recess	cup	food
preparation	food	brother	monkey	skin	eye	video	archive
prejudice	recognition	computer	Internet	money	dollar	dividend	payment
environment	ecology	rook	king	money	currency	drink	mother
association	journal	tiger	feline	journey	car	journey	voyage
shower	thunderstorm	tiger	carnivore	noon	string	stock	market
football	basketball	tiger	mammal	chord	smile	nature	man
morality	importance	tiger	zoo	minister	Party	cell	phone
opera	performance	drink	car	doctor	nurse	lobster	food
chemistry	physics	bishop	rabbi	school	center	boxing	round
psychology	discipline	planet	star	index	benchmark	man	woman
Film	boy	Harvard	Yale	life	death	book	library
cemetery	woodland	planet	moon	life	term	oil	stock
constellation	planet	computation	soccer	gem	jewel	game	series
calculation	cucumber	cup	article	king	cabbage	clothes	clothes
professor	reason	baby	mother	car	flight	line	insurance
hypertension	price	coast	shore	credit	card	man	governor
competition	news	company	stock	vodka	gin	street	avenue
announcement	science	cup	object	drink	ear	cup	liquid
psychology	asylum	planet	galaxy	vodka	brandy	focus	life
madhouse	planet	change	attitude	practice	institution	glass	metal
astronomer	car	football	tennis	viewer	serial	aluminum	metal
automobile	credit	Japanese	American	precedent	law	chance	credibility
information	antecedent	movie	star	space	world	precedent	group
precedent	mind	street	block	admission	ticket	rock	jazz
psychology	memorabilia	sugar	approach	precedent	example	museum	theater
exhibit	issue	problem	challenge	shower	flood	start	match
development	registration	life	lesson	governor	office	summer	drought
arrangement	hike	type	kind	drink	mouth	day	dawn
production	peace	train	car	music	project	ministry	culture

COnférence en Recherche Information et Applications, CORIA'2004.

Harrathi R, Calabretto S., "Une approche de recherche sémantique dans les documents semi-structurés », INFORSID'10, Semantic information retrieval session, Marseille, 2010.

Varelas G., Voutsakis E., Raftopoulou P., Petrakis E. and Milios E., "Semantic similarity methods in WordNet and their application to information retrieval on the web". In: 7th ACM Intern. Workshop on Web Information and Data Management (WIDM 2005), Bremen, Germany (2005) 10-16.

Computing Cross-Lingual Synonym Set Similarity by Using Princeton Annotated Gloss Corpus

Yoshihiko Hayashi

Graduate School of Language and Culture, Osaka University

1-8 Machikaneyama, Toyonaka, 5600043 Japan

hayashi@lang.osaka-u.ac.jp

Abstract

This paper proposes a method to compute cross-lingual semantic similarity between synonym sets. By making use of Princeton Annotated Gloss Corpus as the source of target language statistics, the proposed method exhibited promising results in the experiments: More than 73% of the Princeton WordNet synsets were successfully recovered within the top-5 candidates, given a corresponding set of Japanese WordNet synsets. As the proposed method minimally requires that the input to be seen as an apparently synonymous word set, the method could be extended and the performance would be further improved by incorporating richer information such as textual glosses and/or structural constraints posed by the lexical resources at hand.

1 Introduction

Aligning different knowledge structures, such as ontologies or lexical resources, has long been tackled by several approaches as summarized in (Euzenat and Shvaiko, 2007). The importance of the issue would be much higher, but the methodological direction should be slightly altered, given the recent trends of *Web-servicized* language resources, by which a dynamic yet virtual lexical resource (Calzolari, 2008) could be realized on-demand/on-the-fly. That is, to promptly respond to a user request in such an environment, a robust yet efficient object-to-object *matching* method for finding possible mates in a different resource is more solicited than a structural *alignment*¹, which may require an expensive batch procedure to process the entire resources.

¹An alignment is accomplished as the output of matching processes (Euzenat and Shvaiko, 2007).

With this background, this paper develops a light-weight method to compute cross-lingual semantic similarity, which can contribute to the on-demand/on-the-fly cross-lingual matching of word senses/lexical concepts. The proposed method, which makes use of Princeton Annotated Gloss Corpus² as the source of target language statistics, exhibited promising results in the experiments, demonstrating that the sense-tagged corpus is highly beneficial resource in the presented task.

2 Motivation of the Proposed Method

Lexical semantic resources, in a sense, can be characterized in the light of how the meaning of a word is described and represented. A sense of a word, or a lexicalized concept, could be minimally represented by a set of synonymous words in the same language or in a different language. Such a set of synonyms, a synonym set or a *synset* after Princeton WordNet (PWN), may be further described by the attached textual gloss. As well known, PWN as well as its variant wordnets, including Japanese WordNet (WN-Ja) (Isahara et al., 2008; Bond et al., 2009), model and implement the *relational semantics* (Miller and Fellbaum, 2007). That is, a lexical resource of this type organizes a lexical semantic network, in which each edge carries a lexical semantic relation that holds between the associated lexicalized concept nodes.

In principle, any type of information provided by an actual lexical semantic resource could contribute to the computation of sense similarity. The method proposed in this paper, however, only assumes that the input set of words to be seen as a synonym set. The rationale behind this arrangement is described as follows: (1) Even a poorly annotated lexical resource would provide a kind of synonym set for a sense of the entry word. Some

²It provides a corpus of manually annotated WordNet synset definitions (glosses). It can be download at: <http://wordnet.princeton.edu/glosstag.shtml>.

of the bilingual dictionaries may fall into this category; (2) We could measure a kind of similarity or relatedness between two distinct synonym sets by measuring a cross-lingual *semantic overlap*. To accomplish (2) however, we obviously need to cross the language barrier in some way. To tackle this problem, the presented work makes use of available translation resources backed up by a sense-tagged corpus in the target language, which could play a significant role in narrowing down the potential translation ambiguities.

From a computational perspective, the proposed method can be seen as a light-weight method in the sense that it does not perform a computationally expensive structure-to-structure alignment. If we could achieve a relatively good performance with such a less demanding method, we would be able to develop a far better method on top of it by incorporating additional information, which are often available in lexical semantic resources, into the proposed method. In fact, we may be able to combine the gloss overlap similarity (Lesk, 1986) with the proposed method, again by, for example, translating the source gloss into the target language. Furthermore, we could refine the combined similarity by integrating the extended similarity (Banerjee and Pedersen, 2003; Budanitsky and Hirst, 2006) between the sets of surrounding nodes in the respective semantic networks.

3 Methodology

Given a lexicalized concept s in the source language (SL), the task of the presented work is to acquire a ranked list of possibly corresponding lexicalized concepts $\{t_1, t_2, \dots\}$ in the target language (TL)³. To accomplish this, we compute the semantic similarity between s and t by employing a scoring function $score(s, t)$.

As already discussed in the previous section, the presented method assumes that the input to be seen as an apparently synonymous word set. We thus formulate $score(s, t)$ as follows.

$$score(s, t) = \sum_{x_i \in \sigma(s)} \omega(x_i, s, t) \times score'(x_i, t) \quad (1)$$

Here, $\sigma(s) = \{x_i\}$ denotes the input synonym set; $score'(x_i, t)$ dictates a cross-lingual word-to-concept similarity/relatedness; whereas $\omega(x_i, s, t)$

³More precisely, each of s and t is defined in a lexical semantic resource of the each language.

fundamentally measures the importance or impact of x_i . The weighting function $\omega(x_i, s, t)$ could have various function forms depending on the type of the available lexical resource as discussed in 3.4

3.1 Computing $score'(x_i, t)$ by $p(t|x_i)$

In the presented work, we compute $score'(x_i, t)$ by estimating $p(t|x_i)$, which is the conditional probability of t given x_i . There would obviously be a number of ways to expand $p(t|x)$, but we expand it as follows. Here $\tau(x) = \{y_j\}$ denotes a set of translations in the TL for a word x in the given SL synonym set.

$$p(t|x) = p(t) \sum_{y_j \in \tau(x)} \frac{p(y_j|t)p(y_j|x)}{p(y_j)} \quad (2)$$

The derivation of the equation (2) is detailed in Appendix A, where we applied Bayes' theorem and made the following assumption.

$$p(x|t) \equiv \sum_{y_j} p(y_j|t)p(x|y_j) \quad (3)$$

The rationale behind this assumption is that there may exist a number of hidden elements (TL words); each of which is simultaneously associated with t (by sense relation) and x (by translation relation).

Let us note here that while y_j , in principle, can range over the TL vocabulary, it can be constrained by the condition $p(x|y_j) > 0.0$. In addition, although $p(x|y_j)$ requires TL-to-SL reverse translation probability, this term is removed in the final equation (2), thanks to Bayes' theorem.

Another issue associated with the equation (2) is that a candidate concept t , in principle, can range over the whole set of TL concepts. To be practical, we need to restrict the set of candidate concepts. Two resources can be utilized to accomplish this: (1) The target lexical resource, in our case PWN. We can create a set of candidate synsets by gathering synsets whose synonym set $\sigma(t)$ contains the English word y_j . We call this set S_{syn} ; (2) A sense-tagged corpus, in our case Princeton Annotated Gloss Corpus. We can gather synsets, for which $p(y_j|t)$ is estimated non-zero. We call this set S_{cor} .

In the rest of this section, we describe how $p(y|x)$ and $p(y|t)$ can be computed, and discuss possible formulations of the weighting function $\omega(x_i, s, t)$.

3.2 Computing $p(y|x)$

We constructed two translation probability tables to give $p(y|x)$: (1) $p_d(y|x)$ gives a pseudo translation probability of y given x . We employed more than one bilingual resources provided by EDR Electronic Dictionary (EDR)⁴ (Yokoi, 1995) and assigned pseudo probabilities based on the occurrence number of a word in the set of bilingual resources; (2) $p_c(y|x)$ gives a corpus-based translation probability of y given x . In order to derive this translation probability table, we applied GIZA++ (Och and Ney, 2003), with the ready-made default setting, toward the parallel corpus which comprises of the sense-aligned English glosses (from PWN) and Japanese glosses (from WN-Ja). The gloss texts were morphologically analyzed and lemmatized by using MeCab (Kudo et al., 2004) (for WN-Ja) and TreeTagger (Schmid, 1994) (for PWN). For PWN glosses we also applied stopword deletion and multi-word recognition.

In the following experiments, we also computed $p(y|x)$ by averaging these two probability values. Although the blending ratio could be better determined through experiments, we followed (Xu et al., 2001), which simply calculates the averages.

3.3 Computing $p(y|t)$

Given a sense-tagged corpus, we can readily compute $p(y|t)$ by applying maximum likelihood estimation. We call this $p_c(y|t)$.

Alternatively, the probability value for y that belongs to $\sigma(t)$ can be faked by $p(y|t) = 1/|\sigma(t)|$. We call this $p_d(y|t)$.

In the following experiments, we computed $p(y|t)$ by blending these two probability values with a varying weight α : $(1 - \alpha)p_c(y|t) + \alpha p_d(y|t)$, ($0 \leq \alpha \leq 1$).

3.4 Formulation of $\omega(x_i, s, t)$

It would be natural to utilize a uniform weight, say 1.0, if the words in $\sigma(s)$ are equally important in SL and each $\tau(x_i)$ has almost same impact in determining $p(y|t)$ in TL. As it might not always be the case, we have to devise an appropriate weighting scheme and represent it as a computable function. We considered the following factors in for-

⁴The EDR actually is a dictionary system containing a set of dictionaries organized around the concept system. EDR Japanese-to-English, English-to-Japanese (in a reverse mode), and Concept dictionaries are utilized as bilingual resources.

mulating $\omega(x_i, s, t)$, while we can combine some of these weighting functions to further improve the performance.

- The representativeness of x_i in s . If x is one of the representative words of s , the weight should be higher, while if x is considered as a kind of *noisy* word in s , the weight should be greatly suppressed. To formulate this, we devised $tset(x_i, \sigma(s))$ as detailed below.
- The discriminative power of x_i for s . If x has a skewed association with s , then the weight should be higher, because x can discriminate s among other potentially corresponding SL concepts. We can utilize the inverse document frequency $idf(x)$ to capture this requirement. Note in this case that the document frequency of x dictates the number of lexicalized concepts, in the current lexical resource, whose synonym set includes x as the member. This means that this scheme requires a global figure in the SL lexical resource.
- The *global* similarity between s and t . If s is identified similar to t from a perspective that was not considered in the computation of $p(t|x)$, then $score(s, t)$ would be inflated. To capture this requirement, among several possible measures, we devised $toverlap(s, t)$ as detailed below. Note that this weighting scheme is independent of each x_i in s .

$tset(x_i, \sigma(s))$: This scheme captures the representativeness of x in s by using their translations. Note that these translations have already been acquired while computing $p(t|x)$, hence can be efficiently reused. $Dice(A, B)$ computes the well-known Dice coefficient between set A and set B .

$$tset(x_i, s) = Dice(\tau(x_i), \cup_{x_j \in s \& x_j \neq x_i} \tau(x_j))$$

$toverlap(s, t)$: This scheme simply dictates the cross-lingual concept-to-concept similarity by measuring the overlap between translations of s and the synonym set of t as follows.

$$toverlap(s, t) = Dice(\cup_{x_j \in s} \tau(x_j), \sigma(t))$$

4 Experiments

4.1 Overall Design

To evaluate the performance of the proposed method, we conducted a series of experiments as follows.

In the first two experiments, the task was to *recover*⁵ corresponding PWN synsets given a WN-Ja synset as the query. We first explored an adequate parameter setting for computing $p(t|x)$, while fixing the weighting scheme $\omega(x_i, s, t) = 1.0$ (Experiment-1); we then sought the best weighting scheme $\omega(x_i, s, t)$ that works with the derived $p(t|x)$ configuration (Experiment-2).

The task of Experiment-3, on the other hand, was to *discover* concepts in EDR that could have some association with the given WN-Ja synset. This experiment was conducted in order to assess the applicability of the proposed method to an independently developed lexical resource.

We evaluated the performances by using following measures. As the task, in a sense, is pretty similar to that of information retrieval (IR), we can utilize IR-based measures.

- *t-coverage*: fraction of queries for which at least one English translation was obtained.
- Success rate at rank-n (S@n): fraction of queries for which the target synset was ranked within top-n rank. We consider S@1, S@5, and S@10.
- Mean reciprocal rank (MRR): the average of the reciprocal ranks of the rank at which the target synset was retrieved.

4.2 Experiment-1: Setting Parameters

4.2.1 Data Set

We decided to use Core WordNet⁶ as the source of query concepts to be chosen from WN-Ja. The reasons for this decision was twofold: (1) Core WordNet provides a set of frequent word senses, insisting that a better coverage in translations as well as sense-tag statistics could be expected; (2) Japanese words assigned to the core synsets would be better maintained (Isahara et al., 2008) than the ones assigned to other less frequent synsets.

The query set finally yielded 4,690 synsets (583 adjectives, 3,176 nouns, and 931 verbs⁷), after removing the synsets for which WN-Ja did not assign any Japanese word. We should remind

⁵As each synset in WN-Ja is intrinsically aligned with a synset in PWN, any human-annotated gold-standard data was not required.

⁶<http://wordnetcode.princeton.edu/stand-off-files/core-wordnet.txt>

⁷No adverbs are contained in the Core WordNet.

	t-cov.	S@1	S@5	S@10	MRR
BL-n	0.956	0.072	0.263	0.395	0.177
BL-d	0.943	0.314	0.540	0.661	0.419
BL-c	0.920	0.311	0.409	0.488	0.296
$\alpha = 0.0$		0.365	0.660	0.755	0.497
$\alpha = 0.25$		0.185	0.502	0.675	0.330
$\alpha = 0.5$	0.956	0.180	0.494	0.664	0.324
$\alpha = 0.75$		0.178	0.485	0.654	0.319
$\alpha = 1.0$		0.173	0.477	0.646	0.313

Table 1: The results from the Experiment-1. Each of the bold figures represents the best performance among the comparable conditions. The same applies to the succeeding tables.

again that we did not have to prepare any human-annotated gold-standard data, thanks to the WN-Ja’s pre-aligned organization with PWN.

4.2.2 Results

Table 1 summarizes the results from the Experiment-1, where three baselines (BL-n, BL-d, and BL-c), described below, as well as the results with varying α are shown. Note that we computed a final score by $score(s, t) = \sum_{x_j \in \sigma(s)} p(t|x_j)$, meaning that we applied a uniform weighting scheme.

- BL-n: the set of candidate synsets was created by $S_syn \cup S_cor$, while synset ranking was made only by consulting priors, that is, $score(s, t) = p(t)$.
- BL-d: S_syn was utilized as the set of candidate synsets; $p(y|x)$ and $p(y|t)$ were solely computed by $p_d(y|x)$ and $p_d(y|t)$ respectively. This baseline prescribes a situation in which no sense-tagged corpora are available. Thus, uniform $p(y)$, as well as $p(t)$ were employed.
- BL-c: S_cor was utilized as the set of candidate synsets; $p(y|x)$ and $p(y|t)$ were solely computed by $p_c(y|x)$ and $p_c(y|t)$ respectively. This baseline prescribes a situation in which no translation resources other than the GIZA-aligned translation table are available.
- $\alpha = x$: the set of candidate synsets was same as the BL-n; $p(y|x)$ was computed by averaging the $p_d(y|x)$ and $p_c(y|x)$, as in (Xu et al., 2001), while $p(y|t)$ was computed by $(1 - \alpha)p_c(y|t) + \alpha p_d(y|t)$, ($0 \leq \alpha \leq 1$) to investigate the impact of the resources.

```

Query: 01904930-v
歩く; 歩行; 歩む

* 1 01904930-v 0.980950632623
歩く; 歩行; 歩む; walk
足を使って前進する; 歩いて前進する
use one's feet to advance; advance by steps

2 00283951-n 0.920335762072
ambulation
歩き回ること
walking about

3 10412055-n 0.257621306855
ペデストリアン; 徒歩; pedestrian; footer; walker
徒歩で旅行する人
a person who travels by foot

```

Figure 1: Example of output from the PWN synset recovery task.

As summarized in Table 1, the $\alpha = 0.0$ case outperformed the baselines as well as other cases with $\alpha > 0.0$; the performances monotonically degraded as α increased. In the succeeding experiments, we thus adopted this resource combination and the parameter setting.

We should remark the following: (1) Although the blend ratio might not be optimal, blending two translation resources helped improve the coverage as well as the performance. (2) As the result with $\alpha = 0$ indicates, the statistics from the sense-tagged corpus largely outperformed the faked statistics originated from the bilingual dictionaries. (3) Among the baselines, BL-d performed relatively well in all measures, while BL-c exhibited only good in S@1, suggesting that the coverage in the sense-tagged corpus statistics was not broad enough nor uniform. This could be quite reasonable, because the body of the corpus was glosses taken from the PWN synsets.

4.3 Experiment-2: Recovering PWN Synsets

To illustrate how the output of the method looks like, Figure 1 exemplifies the output, in which the query synset (id:01904930-v) with the Japanese synonym set {歩く, 歩行, 歩む} is correctly ranked first among others. Notice that we did not apply part-of-speech constraint, as the ranked second and third noun synset show.

4.3.1 Results

Table 2 sums up the results. Although not shown in the table, t -coverage was 0.956 for all cases. The bottom row of the table dictates a kind of upper-limit case, which is defined as: $sset(x_j, \sigma'(t)) \times tooverlap(s, t)$. Notice that we can never utilize this scheme in an actual task of

Weighting scheme	S@1	S@5	S@10	MRR
Baseline	0.365	0.660	0.755	0.497
idf	0.365	0.659	0.757	0.498
tset	0.474	0.716	0.790	0.582
tooverlap	0.397	0.687	0.778	0.527
idf \times tooverlap	0.398	0.688	0.778	0.528
idf \times tset	0.475	0.717	0.791	0.583
idf \times tset \times tooverlap	0.493	0.737	0.813	0.603
Upper-limit	0.634	0.915	0.948	0.754

Table 2: The results from the Experiment-2.

mate finding across language resources, because it incorporates $sset(x_j, \sigma'(t))$, which is defined as $sset(x_j, \sigma'(t)) = 1/|\sigma'(t)|, (x_j \in \sigma'(t))$. Here, $\sigma'(t)$ denotes a set of Japanese words associated with the PWN synset t , meaning that this scheme *illegally* utilizes the information that can only be acquired from WN-ja, which, needless to say, has already been aligned with PWN.

The figures shown in Table 2 are reasonably promising, given the fact that PWN synsets are often considered overly fine-grained (Miller and Fellbaum, 2007). From Table 2, we can safely say: (1) The discriminative power given by $idf(x)$ slightly contributed to the performance improvement, while $tset(x, s)$, capturing the representativeness of a set element, greatly improved the performance. By closely looked at the weights given by $tset(x, s)$, we noticed that assigning zeros to potentially irrelevant synsets were quite effective. That is, generating a richer translation set first is necessary, but being backed up by a proper filtering mechanism is crucially important. (2) By combining these two weights, the MRR measure was improved from 0.497, achieved by the baseline, to 0.583, showing more than 17% improvement. (3) The performance was further improved to 0.603 (21.3% improvement) by incorporating the global weight dictated by $tooverlap(s, t)$.

While the results obtained from this experiment were promising, we should and would be able to further improve the weighting scheme, because there still exist performance gaps between the best-so-far case and the upper-limit case.

To help drill down the potential issues, Table 3 breaks down the results from the best-so-far case and the upper-limit case by the part-of-speech of the query synsets. The parenthesized numbers are ratios to the corresponding figures brought about by the upper-limit case. One of the tendencies obvious in the table is noun concepts are far easier to identify than adjective and verb concepts. In par-

Weighting scheme	POS	S@1 (%)	S@5 (%)	S@10 (%)	MRR (%)
idf × tset × tooverlap	adj	0.350 (67.0)	0.563 (70.3)	0.658 (80.0)	0.452 (70.6)
	noun	0.569 (81.9)	0.809 (85.1)	0.869 (89.1)	0.676 (84.1)
	verb	0.310 (63.3)	0.586 (68.7)	0.705 (76.5)	0.439 (67.3)
	total	0.560 (88.3)	0.761 (83.2)	0.829 (87.4)	0.652 (86.5)
Upper-limit	adj	0.522	0.801	0.822	0.640
	noun	0.695	0.951	0.975	0.804
	verb	0.490	0.853	0.922	0.644
	total	0.634	0.915	0.948	0.754

Table 3: Part-Of-Speech breakdown of the Experiment-2 results.

Relevance level	Weighting scheme	S@1	S@5	S@10	MRR
Syn	Baseline	0.412	0.663	0.724	0.522
	idf	0.432	0.683	0.714	0.540
	idf × tset	0.441	0.676	0.706	0.546
	idf × tset × tooverlap	0.412	0.689	0.724	0.533
	Upper-limit	0.427	0.709	0.749	0.543
Rel	Baseline	0.683	0.859	0.910	0.763
	idf	0.688	0.859	0.910	0.765
	idf × tset	0.719	0.884	0.915	0.792
	idf × tset × tooverlap	0.724	0.890	0.915	0.796
	Upper-limit	0.714	0.920	0.930	0.799

Table 4: The results from the Experiment-3.

```

Query: 0eced1
Jsynonyms:弱腰; 憶病さ; 慢病; 不覚悟; 小心; 脳病; 気弱
Esynonyms:cowardice;timidness;faint-heartedness;
timidity;poltroonery;fearlessness;recrancy;timorousness
Jgloss:気が弱いこと
Egloss:the condition of timidity

** 1 04860065-n 0.0271817807978
腰抜; 物怖じ; 腹抜け; 怖懦; 小胆; おく病; 慢病; 小胆さ; 慢病さ; 脳病さ; 小心; 脳病; 腹ぬけ; 腹抜け; cowardice; cowardliness
勇気における特性
the trait of lacking courage

* 2 04860759-n 0.00644116122449
内氣; おく病; 慢病; 脳病; 気弱; timidity; timorousness
新しく、未知の場所や活動に思い切って乗り出すときの恐怖心
fearfulness in venturing into new and unknown places
or activities

** 3 00264776-a 0.00429356521739
ふがいない; 小気; 怯弱; 意気地のない; 脇甲斐無い; 意気地無い; 柔弱; 心弱い; だらしない; 軟弱; 温い; だらし無い; 脇甲斐ない; 小胆; 慢病; 不甲斐無い; 女々しい; 小心; 脳病; 意気地ない; 不甲斐ない; 心よいわい; cowardly; fearful
勇気が欠如しているさま; 卑劣に脳病で気が弱い
lacking courage; ignobly timid and faint-hearted

```

Figure 2: Example of output from the PWN synset discovery task (**:Syn-level, *:Rel-level).

ticular, to improve the S@1 figure for verb concepts seems quite difficult. Although it is readily expected that the degree of word sense ambiguity defined in PWN affected the results, we should carefully look through each of the defeated cases.

4.4 Experiment-3: Discovering Mate Synsets for Concepts in EDR

Figure 2 exemplifies the output, in which possibly corresponding PWN synsets are discovered

against the EDR query concept (id:0eced1).

Similar to PWN synsets, many of the concepts in EDR are associated with a set of Japanese and/or English words that participate in some of the sub-dictionaries of EDR. Also many of the concepts are glossed in Japanese and/or English; and the concept nodes form a kind of semantic network by being connected each other with conceptual/semantic relations. In this sense, at least from the perspective of information structure, EDR can be modeled as a WordNet-type semantic lexicon (Savas et al., 2010). However unlike PWN, any EDR concept is not classified by the part-of-speech of the associated words, meaning that an EDR concept is POS-independent.

4.4.1 Data Set

Unlike the previous experiment of PWN synset recovery, we had to prepare an evaluation data set for this experiment, because no innate gold standards exist for any EDR concept. This time we constructed a small-sized evaluation data set comprises of 203 query concepts and the assigned relevance judgments through the following procedure.

1. Choosing query concepts: Among more than 400,000 EDR concepts, we applied the following conditions, and then randomly chose 203 concepts: (1) A concept has to have one or more Japanese synonym words; (2) It has to have both Japanese and English glosses;

(3) At least one of the senses of Japanese synonyms have to be considered familiar by Japanese native speakers. This may be a special requirement, but it is introduced so that the query set would be similar to Core WordNet, in the sense it gathers frequent and familiar concepts. To this end, we consulted a Japanese lexical resource called Lexeed (Fujita and Nagata, 2010) in which a word sense familiarity figure is assigned to the word sense entries.

2. Collecting candidate PWN synsets: By applying one of the already described methods ($\alpha = 0.0$, uniform weight), we collected at most 25 candidate synsets for each query concept. Collected ranked lists were then given to a human annotator.
3. Annotating relevance judgments: One of the following labels was assigned to each candidate synset by a human annotator if it is not irrelevant. The annotator, not the author of this paper; a Japanese native speaker with a relatively high English literacy, was requested to annotate just by referring to EDR synonyms/glosses given in Japanese and English and PWN synonyms/glosses given in English.
 - Almost-synonymous (Syn-level): A synset considered almost synonymous is given this label. As a concept in EDR is not classified by part-of-speech, this label was also assigned to POS variants of a thought-to-be relevant concept. This led the cases where one or more candidate synsets were assigned this label. In the evaluation however, only the highest ranked candidate was considered correct. Also, not always this label was given to any synset in the candidate set, due to the sense gap between EDR and PWN concepts and/or missing of relevant synsets in the collected candidate set.
 - Related-somehow (Rel-level): A synset that may have some association with the EDR query concept is assigned this label. The association could be: hyponymy, meronymy, agentive argument or even more informal relationships. It is natural that more than one candidate synsets were assigned this label. As in the Syn-level, only the highest ranked

candidate was considered in the utilized measures.

4.4.2 Results

Table 4 summarizes the results, where *t-coverage* was 0.980 for all cases. Remind that a Rel-level figure contains the number of Syn-level matching. As the best weighting cases show, S@5 in Syn-level approaches to 70%, while that of in Rel-level approximates 90%, insisting that the proposed method could be effectively utilized as a tool to associate two distinct lexical resources.

Points that should be made from this table are threefold: (1) The scheme mentioned as upper-limit does not function as upper-limit any more. This is quite reasonable, because the query synonyms are coming from EDR, which is a totally different resource from PWN; (2) In Syn-level, the effect of *tset* was significant in improving S@10 and MRR, while combining it with *toverlap* improved S@5 and S@10. We should further investigate the results, but it can be said that the Japanese synonym sets in EDR were relatively uniform and less noisy compared to that of WN-Ja; (3) In Rel-level, although the baseline exhibited a steady performance, *tset* combined with *toverlap* outperformed it. Interestingly the combined method also surpassed *tset*. At the moment the reason is unclear, but this should be further examined, because it might provide some clue to distinguish synonymy identified in the Syn-level with other lexical semantic relations identified in the Rel-level.

5 Concluding Remarks

The proposed method, which only requires a set of apparently synonymous SL words as the input, exhibited promising results in computing cross-lingual semantic similarity between synonym sets. Encouraged by the results, we are now, in expectation of gaining substantially higher performances, investigating incorporation of other types of information, such as textual glosses and structural constraints available in lexical semantic resources.

Comparing our method with a representative method such as proposed in (Agirre et al., 2009), ours accepts a set of synonymous SL words as the input, rather than a single word. To enable this, weighting schemes to integrate word-to-synset similarities were devised and evaluated. However a more robust scheme could be further explored by evaluating it with a larger data set. In

the course of the work, we would also devise a computational method to distinguish between the synonymous relation and other lexical semantic relations.

The proposed method obviously benefited from the use of the Princeton Annotated Gloss Corpus as the source of TL statistics, particularly in handling frequent and familiar concepts. This means that if we are to broaden the coverage to less frequent concepts, we need to have larger sense-tagged corpora or explore alternative resources good for dealing with translation ambiguities.

Acknowledgments

The author would like to thank Chiharu Narawa for the helpful discussions. The presented work was supported by KAKENHI (21520401) provided by MEXT, Japan, and partly by the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the MIC, Japan.

References

- Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pașca, and Aitor Soroa. 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. *NAACL-HLT 2009*, pp.19–27.
- Satyanjeev Banerjee and Ted Pedersen. 2003. Extended Gloss Overlaps as a Measure of Semantic Relatedness. *Proc. IJCAI 2003*, pp.805–810.
- Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kurabayashi, and Kyoko Kanazaki. 2009. Enhancing the Japanese WordNet. *Proc. the 7th Workshop on Asian Language Resources*.
- Nicoletta Calzolari. 2008. Approaches towards a 'Lexical Web': the Role of Interoperability. *Proc. ICGL 2008*, pp.34–42.
- Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, Vol.32, No.1, pp.13–47.
- Jérôme Euzenat and Pavel Shvaiko. 2007. *Ontology Matching*. Springer.
- Sanae Fujita and Masaaki Nagata. 2010. Enriching Dictionaries with Images from the Internet. - Targeting Wikipedia and a Japanese Semantic Lexicon: Lexeed -. *Proc. COLING 2010*, pp.331–339.
- Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanazaki. 2008. Development of the Japanese WordNet. *Proc. LREC 2008*, pp.2420–2423.
- Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. *Proc. EMNLP 2004*, pp.230–237.
- Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a pine cone from an ice cream cone. *Proc. SIGDOC 86*, pp.24–26.
- George A. Miller and Christiane Fellbaum. 2007. WordNet Then and Now. *Language Resources and Evaluation*, Vol.41, pp.209–214.
- Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, Vol.29, No.1, pp.19–51.
- Bora Savas, Yoshihiko Hayashi, Monica Monachini, Claudia Soria, and Nicoletta Calzolari. 2010. An LMF-based Web Service for Accessing WordNet-type Semantic Lexicons. *Proc. LREC 2010*, pp.507–513.
- Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proc. International Conference on New Methods in Language Processing*, pp.44–49.
- Jinxi Xu, Ralph Weischelde, and Chanh Nguyen. 2001. Evaluating a Probabilistic Model for Cross-lingual Information Retrieval. *Proc. SIGIR 2001*, pp.105–110.
- Toshio Yokoi. 1995. The EDR Electronic Dictionary. *Communications of the ACM*, Volume 38, Issue 11, pp. 42–44.

Appendix A. Derivation of $p(t|x)$

By Bayes' theorem, $p(t|x)$ is rewritten as:

$$p(t|x) = \frac{p(t)p(x|t)}{p(x)} \quad (4)$$

Here, we assume the equation (3). This equation is expanded as follows, again by applying Bayes' theorem.

$$\begin{aligned} p(x|t) &\equiv \sum_{y_j} p(y_j|t)p(x|y_j) \\ &= \sum_{y_j} p(y_j|t) \frac{p(y_j|x)p(x)}{p(y_j)} \end{aligned}$$

By plugging this into (4) and adjusting the summation range, we get the equation (2).

$$p(t|x) = p(t) \sum_{y_j \in \tau(x)} \frac{p(y_j|t)p(y_j|x)}{p(y_j)} \quad (2)$$

Restructuring Adjectives in WordNet with *ClusterEditor*

Isaac Julien

Princeton University

ijulien@princeton.edu

Christiane Fellbaum

Princeton University

fellbaum@princeton.edu

Abstract

We discuss the shortcomings of the current representation of adjectives in the Princeton WordNet and present a proposal for a comprehensive reorganization. First, a clustering algorithm applied to Web data allows us to determine the adjectives' compatibility and to (re)group senses accordingly. Next, we discuss the integration of adjectives into a relational database that is currently being developed for WordNet. To support this work, we developed an editor and a GUI that allow users to assign adjectives to the semantically appropriate groups. The clustering algorithm is adapted specifically to work with the editor.

1 Introduction

WordNet currently organizes ascriptive adjectives into “dumbbell” structures (Gross and Miller 1990, Miller 1998). Two antonymous head synsets (in most cases singlets) are connected by an arc expressing lexical-semantic antonymy, and each head synset is linked to one or more “satellite” adjective synsets by a link denoting “semantic similarity.” Additionally, each head synset may be linked to a noun attribute that denotes the property shared by all adjectives.

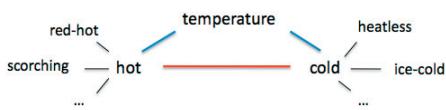


Figure 1: WordNet Dumbbell Structure

For example, in Figure 1, the antonymous heads (also called “centroids”) are “hot” and “cold.” “Hot” is linked to similar adjectives “red-hot” and “scorching,” and “cold” to “heatless” and “ice-cold.” We refer to the structure comprising the centroid and its satellites as a “cluster.” All adjectives in the dumbbell express specific values of the attribute “temperature.” While the representation of adjectives in WordNet differs significantly from that of nouns and verbs, it is founded on semantic similarity, a cornerstone of WordNet that has proved to be beneficial for applications requiring word sense disambiguation.

This organization of WordNet adjectives has several benefits. First, as described by Gross, Fischer and Miller (1989), it captures the strong mutual association between the antonymous centroids. Justeson and Katz (1991) demonstrated that centroids also show particular distributional properties not shared by their satellites or other adjectives. Second, the representation of a cluster in terms of a centroid and satellite allows for the subtle distinction between “direct” (centroid-centroid) and “indirect” (satellite-centroid and satellite-satellite) antonyms across clusters; this distinction, too, receives support from psycholinguistics, the adjectives’ frequencies and distributional properties. Third, the “attribute” link is one of the few arcs in WordNet connecting words and synsets from different parts of speech. The current sparsity of cross-POS links is a weakness of WordNet’s design and limits its usefulness for NLP applications.

Despite its appealing design, the dumbbell structure has some significant problems, some of which we attempt to address here.

2 Problems with the Dumbbell Structure

We identify several shortcomings of the current organization of adjectives in WordNet. The first and most significant is the vague and heterogeneous “similar” link, which does not capture a specific, consistent relation between either the satellites and the centroid or among the satellites. As a result, some adjectives that are currently assigned to the same cluster do not form a semantically coherent group.

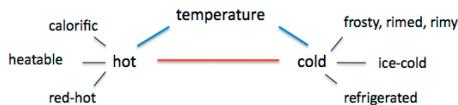


Figure 2: Temperature Adjective Cluster

Figure 2 shows part of the cluster of adjectives expressing values of the attribute “temperature.” The centroid “cold” is similar to the synsets {frosty, rimed, rimey}, {ice-cold} and {refrigerated}. But while “ice-cold” expresses a value of the attribute “temperature,” “frosty” and “refrigerated” arguably do not. Something that is frosty or refrigerated might well be inferred to be cold, but the adjectives themselves do not describe temperature: “frosty” and “rimed” relate to frost, but not to temperature per se; “refrigerated” means no more than that an entity is placed inside a refrigerator, and while it may have cooled there, this does not necessarily follow. Similarly, “calorific” is classified as being similar to “hot,” even though “calorific” does not express a degree of temperature but instead refers to the ability to generate heat (WordNet defines “calorific” as “heat-generating”). “Heatable” simply refers to the property of an entity that can be heated; thus, water is inherently heatable but it is not always warm or hot.

Moreover, similar adjectives in the same cluster often evaluate different senses of the cluster’s attribute. For example, adjectives “bulky” and “epic” are classified as being “similar” and belonging to a cluster related to the attribute “size.” However, while “bulky” describes physical size, “epic” does not and it

should be assigned to another cluster of adjectives expressing values of non-physical size.

The intuitively obvious lack of semantic homogeneity among adjectives in a cluster is reflected in their different selectional restrictions, i.e., the fact that they modify distinct classes of nouns. In Figure 2, both “cold” and “ice-cold” can describe anything that has a value for temperature (“this food is ice-cold,” “the weather is cold”; “his feet were cold,” etc.) but “refrigerated” is far more restricted (“this food is refrigerated,” **the weather is refrigerated**, **his feet were refrigerated**).

The infelicitous assignment of “similar” adjectives to a centroid results in inaccurate pairings of indirect antonyms, such as “calorific-cold.” Similarly, the indirect antonym of “refrigerated” is “hot,” whereas “unrefrigerated” or “heated” would be more appropriate. Indirect antonyms are characterized by their relatively low frequency and tighter selectional restrictions compared to those of the centroids. In most cases, the centroid selects for a wider class of nouns than its satellites. This is a consequence of the centroid being semantically underspecified, whereas the meanings of the satellites are more specific and hence apply to a narrower class of nouns. (This, in turn, accounts at least in part for the lower frequency of satellites.) Thus, many of the adjectives currently included in a single cluster may be candidates for assignment to different clusters whose centroids share the same word form but express different senses. Such a representation would pair adjectives that are similar (or antonymous) by virtue of shared selectional restrictions. Finally, “attribute” links have not been consistently been encoded. Many clusters that intuitively seem to share an obvious attribute are not connected to one. For example, the centroids “rich” and “poor” should probably be linked to the attribute “wealth.” (We leave aside here the question concerning the overlap of attributes and nouns linked to adjectives via “morphosemantic” links as well as “pertainingms”.)

3 A Suggested New Structure

We propose a revised structure for the adjectives in WordNet that maintains the attribute and antonymy links but does away with the

undefined “similar” are among centroids and satellite synsets. Instead of clusters containing a centroid linked to similar adjectives, the proposed structure consists of smaller groups of adjectives that are indeed semantically similar in that they express values of the same attribute (in many cases, a particular sense of a polysemous noun). Each group (or cluster, a term we will retain) of adjectives is linked to an attribute, and may be linked to an antonymous group of adjectives linked to the same attribute. No one adjective in each group is singled out as a centroid. However, the original antonymy link between centroids - a lexical relation between two words, not between synsets or clusters - will be preserved.

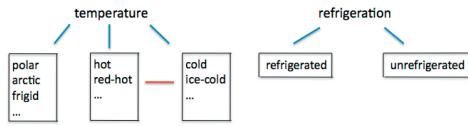


Figure 3: Suggested New Structure

Figure 3 shows an example of the proposed new structure. Among the adjectives expressing values of the attribute “temperature,” the cluster [“polar”, “arctic,” “frigid”] is distinct from [“cold,” “ice-cold”] because the adjectives in the former typically modify the weather, whereas “cold” and “ice-cold” additionally can describe physical objects. As such, they are antonyms of “hot” and “red-hot,” as indicated by the red line in the figure. (Senses of “hot” and “cold” that do not describe temperature are in different clusters and relate to different attributes.)

The revised structure preserves the current semantic information about antonymy but it is more precise in that it includes only strictly related adjectives. We have described elsewhere (Sheinman et al., submitted) in detail specific proposals for the reorganization of scalar adjectives in WordNet, based on Sheinman and Tokunaga (2009). In the remainder of this paper, we focus on the description of a new editor that will facilitate the manual restructuring of WordNet’s adjective component along the lines outlined above. The tool will be made freely available so as to allow builders of wordnets in other languages to create or edit adjectives. A unified approach to the representation of adjectives across languages will significantly

improve crosslingual comparisons and NLP applications.

3.1 ClusterEditor

To restructure WordNet adjectives as described in the previous section, a tool is needed that will allow the manual reassignment of adjectives in the current clusters to new clusters. ClusterEditor is a java GUI designed to make the restructuring process as fast, easy, flexible, and accurate as possible.

ClusterEditor allows a user to view and edit adjective clusters in the current WordNet database. Its main functionality is to allow the reformatting of these clusters into the proposed new ones.

3.2 WordNet as an SQLite Database

ClusterEditor is designed to modify the WordNet database in a relational (SQLite) format. A number of users have proposed relational databases for WordNet. Bernard Bou makes a WordNet SQLite database freely available online with his *WordNetScope* project (Bou, *WordNetScope*). Ernst (2010) discusses in detail the significant advantages of a relational format over the current limited text-based one. Working with an SQLite database moreover makes restructuring the adjectives a much more approachable task. In anticipation of the transition to a relational database, we work with an SQLite version of WordNet.

3.3 Modifying the Current Database

For full documentation on ClusterEditor, see *ClusterEditor Documentation* Available on the WordNet site at <http://wordnet.princeton.edu/>.

The main window of ClusterEditor allows a user to view WordNet adjectives in their current structure. ClusterEditor displays a list of noun attributes; the user can select an attribute to see the adjective centroids linked to it, and can then select a centroid to see its “similar” (satellite) adjectives. For the many clusters that are not linked to an attribute, an option to view them directly is available.

ClusterEditor allows the user to make several kinds of changes to the current structure.

First, the definition (gloss) of a selected centroid or satellite adjective can be edited. This allows for more precise and specific descriptions of the meanings of adjectives. Importantly, many algorithms based on WordNet incorporate the Lesk method (Lesk 1986), which uses the lexical overlap among definitions, and crisp and clearly distinguishable definitions may well improve results. Second, ClusterEditor allows users to move satellite adjectives to new centroids, and to move centroids to different attributes. In this way, the user may reassign existing adjectives without a complete restructuring and without adding or deleting adjectives. There many inconsistencies among the “attribute” and “similar” links make this a necessary feature. As mentioned before, many clusters are currently unconnected to the nouns that express the relevant attributes. Any changes made to the current adjective structure are recorded in a file “log.txt” created in the same directory as ClusterEditor and can be reviewed or undone.

3.4 Restructuring Adjectives

The main functionality of ClusterEditor is to reformat the current WordNet adjective clusters into the new clusters proposed above. The user can select one or more centroids to reformat. Some currently separate clusters like “hot” and “warm” are obvious candidates for grouping. The reformatting window displays a list of all the adjectives chosen for reformatting and allows the user to remove those that do not belong in a given cluster. The user can then either manually divide the adjectives up into new clusters, or use the automatic clustering tool described in the next section. The user can also select attributes for each of the clusters; ClusterEditor will suggest a list of attributes to choose from based on words in each cluster, or the user can search for one in a popup window.

When the user finalizes a new cluster to be added to the database, ClusterEditor checks if any other clusters already exist under the attributes selected. In that case, the user has the option to merge the new cluster(s) with the existing ones or to represent them as antonyms. This prevents the assignment of related adjectives to different, unrelated clusters.

3.4 New Database Tables

ClusterEditor creates three new tables in the SQLite database: “clusters,” “attributes,” and “antonyms.” The “clusters” table contains synsetids of adjectives as the primary key and clusterids, randomly generated integers between 1 and 999,999,999. Adjective synsets in this way are connected to the appropriate clusters. The “attributes” table links clusters to attributes (e.g., adjectives like *large*, *big* and *huge* are assigned to the attribute *size*), and contains the clusterid (primary key) and synsetid of the attribute. The “antonyms” table links antonymous clusters and contains the clusterids of the two clusters (the first clusterid is the primary key).

The implication of using the synsetid of adjectives as a primary key is that no adjective can belong to more than one cluster. While this constraint may seem overly strict, we argue that if an adjective synset appears to belong to two clusters, it has two distinct senses and should be split into two different synsets accordingly. This principle of unique form-meaning mappings underlies all of WordNet.

4 Clustering Adjectives

We explore ways to separate WordNet adjective clusters automatically into groups of closely related adjectives. The goal is to facilitate the reorganization of adjectives into the proposed new structure and reduce the considerable manual effort involved in this task.

Given that selectional restrictions serve to distinguish meanings, we first examine the types of nouns that each adjective in a cluster can modify. For example, in the cluster in Figure 2, we might want to separate “refrigerated” from “cold” and “ice-cold.” This move would be supported if we found that set of nouns modified by “cold” is similar to that modified by “ice-cold,” but different from the set of nouns described as “refrigerated.”

We search the OANC corpus [OANC], part-of-speech tagged with the ANC Tool [OANC] for syntactic constructions including an adjective and a noun that it modifies, such as “[adjective] [noun]” and “[noun] is/are/was/were [adjective]” (“an expensive watch,” “this watch is expensive”). The OANC is a balanced and contemporary corpus, and we expect the extracted data to be representative.

4.1 Applying a Distance Measure

Next we define a semantic distance measure among adjectives, based on the sets of nouns found in the OANC that are modified by the adjectives. We calculate the shortest path to the lowest common superordinate (the Shortest Ancestral Path, or SAP, Ene and Wayne, 2006; Resnik 1995, Leacock and Chodorow 1998, *inter alia*) between every noun described by the first adjective and every noun described by the second.

4.2 The Clustering Algorithm

The clustering algorithm takes a group of adjectives from one of the WordNet dumbbells as input. After identifying the nouns modified by each adjective and computing the distance between each pair of adjectives, as described above, the clustering algorithm splits up the old cluster into one or more new clusters.

The adjectives are initially in a single group. The algorithm calculates the distance of each adjective from the group as the average distance to all of the other adjectives in the group. The distance of each adjective to the group is compared to the average of these distances. If it is greater than one standard deviation above the average, the adjective is removed from the group. When an adjective is removed from a group, it is either placed in a new group or moved to an existing group. The algorithm adds it to an existing group if its average distance to the members of that group is less than one standard deviation above the average distance of the members of that group to the group. This process repeats for each group until either the algorithm stops making changes between iterations or reaches a preset limit on the number of iterations (currently set to 1000). The resulting groups are the new clusters. We initially used k-means clustering, starting with a number of clusters that depended on the number of adjectives being clustered, and merging clusters whose members were close enough to each other. However, the results varied too much from run to run, and we abandoned this approach in favor of the current method.

4.3 A Sample Result

For the attribute “weight,” the algorithm produced the clusters in Figure 4 below. Notice that the synset [“doughy,” “soggy”], which does not describe weight, has been separated out, and that [“dense”] and [“massive”], which describe mass, have been grouped together apart from [“heavy”].

Cluster 1: [dense], [massive]

Cluster 2: [heavy], [hefty], [ponderous]

Cluster 3: [doughy, soggy]

Figure 4: Sample Result of Clustering

5 Limitations and Remaining Problems

The OANC corpus, which we chose for its currency, balancedness, and POS tags, with roughly 15 million words of spoken and written dialogue [OANC], is however too small to include many of the less common adjectives in WordNet clusters or to include a reliable number of tokens. The clustering algorithm requires some minimum number of nouns for each adjective, and cannot accurately cluster these uncommon adjectives. Second, the algorithm makes no distinction between different senses of an adjective, since the corpus is not sense-tagged. For uncommon senses of an adjective with a different predominant sense (“cold” in the sense of “emotionless”), the accuracy of the clustering is thrown off.

Finally, there is a trade-off between the time the clustering algorithm takes to run and the accuracy of the results. Considering more nouns for each adjective means performing more SAP calculations, the most expensive part of the algorithm. Considering fewer nouns means a higher risk of ending up with erroneous results.

6 Adapting the Algorithm for ClusterEditor

With these limitations in mind, we adapt the clustering algorithm for use in ClusterEditor. First, we divide the algorithm into two steps: the initial data gathering where the distances between all pairs of adjectives is calculated, and the actual clustering. The first step is the time consuming one, so this allows the user, after clustering once, to adjust the sensitivity of the

algorithm and cluster again almost instantly. Second, we allow the user to make changes to the groups that result from clustering. In that sense ClusterEditor uses the clustering algorithm to provide a suggestion of what the new clusters should be. The user can then adjust the sensitivity of the clustering and move or delete adjectives from the clusters. This adds flexibility to the reformatting process, and since other steps such as attribute selection must be done manually, there is not much benefit to clustering completely automatically.

Another question concerns synsets with more than one lemma. While the reorganization must be done on the level of synsets, the pattern searching for the clustering algorithm is done for each lemma in a synset, so lemmas that initially were assigned one same synset may end up different synsets and even distinct clusters. To place synsets into clusters, the algorithm chooses the cluster that the majority of its lemmas ended up in.

6 The Database GUI

When working with ClusterEditor, a user might want to view and edit the new structures. DBGUI is another java GUI that allows flexible control over reorganized clusters and their antonyms and attributes. For full documentation on DBGUI, see *DBGUI documentation* (Available on the WordNet site at <http://wordnet.princeton.edu/>).

6.1 Modifying the Adjective Representation

DBGUI allows the user to view all of the adjective clusters that have been added to the new database tables, organized by attribute. It displays the members of each cluster, as well as any antonymous clusters. The user can make any necessary changes by adding or removing adjectives from clusters, deleting clusters, and adding or removing antonyms. If ClusterEditor produces clusters that the user later wants to check on or change, DBGUI makes it easy to do so.

7 Conclusion

We addressed some of the major problems with the current representation of adjectives in WordNet. Arguing for an improved representation that would result in semantically

clean clusters, we describe an editing tool that allows the manual reassignment of adjectives in WordNet to different clusters and to the corresponding attribute nouns. We propose to harness corpus data that will identify semantically (dis)similar adjectives on the basis of their selectional restrictions. It is hoped that the work described here will lead to a renewed interest in adjectives and that the potential of ClusterEditor will be harnessed by developers of wordnets, leading to improved results in NLP applications.

Acknowledgment

Julien's and Fellbaum's work was supported by grant No. CNS 0855157 from the U.S. National Science Foundation and by the Tim Gill Foundation.

References

- Bernard Bou, *WordNetScope*.
<http://wnscope.sourceforge.net/>
- Alina Ene and Kevin Wayne. 2006. *WordNet Programming Assignment* for course COS 226, Department of Computer Science, Princeton University.
- Adam Ernst. 2010. *A Relational Version of WordNet with Evocation Ratings and its Application to Word Sense Disambiguation*. Senior Thesis, Computer Science Department, Princeton University.
- Derek Gross, Ute Fischer and George A. Miller. 1989. The organization of adjective meanings. *Journal of memory and Language*, 28, 92-106.
- Derek Gross and Katherine J. Miller. 1990. Adjectives in WordNet. *International Journal of Lexicography* 3 (4): 265-277.
- Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In: *WordNet: An Electronic Lexical Database*, Ed. C. Fellbaum. Cambridge, MA: MIT Press, 265–283.
- Lesk, Michael. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: *Proceeding of the 5th annual international conference on Systems documentation (SIGDOC)*.

Miller, Katherine J. 1998. Modifiers in WordNet. In: *WordNet: An Electronic Lexical Database*, Ed. C. Fellbaum. Cambridge, MA: MIT Press.

Open American National Corpus (OANC)
<http://americannationalcorpus.org/OANC/index.html>.

Philip Resnik. 1995. Using information content to evaluate semantic similarity. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 448–453, Montreal.

Vera Sheinman, Christiane Fellbaum, Isaac Julien, Peter Schulam and Takenobu Tokunaga. Submitted. Large, Huge, or Gigantic? Identifying and encoding intensity relation among adjectives in WordNet. *Language Resources and Evaluation: Special issue on wordnets and relations*.

An Extractive Approach of Text Summarization of Assamese using WordNet

Chandan Kalita
Department of CSE
Tezpur University
Napaam, Assam-784028
chandan_kalita@yahoo.co.in

Navanath Saharia
Department of CSE
Tezpur University
Napaam, Assam-784028
nava_tu@tezu.ernet.in

Utpal Sharma
Department of CSE
Tezpur University
Napaam, Assam-784028
utpal@tezu.ernet.in

Abstract

Automatic text summarization means finding out the summary of one or more document by a computer program. The output text or the summary should contain the most important points of the original text without changing its meaning. In this report, we present an extractive approach of Text summarization of Assamese, a free word order inflectional Indic language, using WordNet. From our experiment, we got approximately 78% accurate result.

1 Introduction

Automatic text summarization means finding out the summary of one or more document by a computer program. The output text or the summary should contain the most important points of the original text without changing its meaning. With huge amount of information available on the World Wide Web, there is a pressing need to have “Information Access” systems that would help users by providing the relevant information in a concise, pertinent format. Two main category of text summarization are (Das and Martins, 2007).

- Extractive
- Abstractive

Extractive approach is a procedure of identifying important sections of the text and reproducing them as they are. There are no modifications done in the input text pattern. In this approach, there are mainly two steps- *Extraction* and *Fusion*. In the extraction step, important sections are identified and extracted sections are combined coherently in the fusion phase. In abstractive approach considerable amount of linguistic

analysis is performed for the task of summarization. In this approach, important sections of the input text are identified and produced in a new way. In the abstractive approach new sentence are generated without changing the topic meaning. In this paper, we present certain aspects of “Text Summarization” and implement one extractive approach for Assamese language that is the easternmost Indo-European language with around 30 million speakers.

In the next section, we describe prior works in single and multi-document text summarization. Section 3 and 4 describe preprocessing phases: similarity measures used in our approach for summarization of Assamese text and obtained results respectively. Section 5 concludes our paper.

2 Literature Survey

2.1 Single document summarization

In the 1990s, with the advent of machine learning technique used in NLP, a series of seminal publications appeared that employed statistical techniques to produce document extracts.

Naive-Bayes Method:

In this method the program is able to learn from existing data. A classification function determines for each sentence whether the sentence should be included in the summary or not using a Naive Bayes classifier. In this approach a score is given to each sentence and only the n top sentences are extracted.

Rich Features based Method:

Lin and Hovy, 1997 studied the importance of a single feature- “sentence position”. Just weighing a sentence by its position in the text, which the authors term as the “position method”, is based

on the idea that texts generally follow a predictable discourse structure, and the sentences of the main topic tend to occur in certain predefined locations (e.g. title, abstracts, etc). However, since the discourse structure significantly varies over domains, the position method is not a good choice. Naive-Bayes method and Rich Features based methods are some example of sentence extraction based summarization approach. There are some other approaches whose working principles are different but use extraction based. For example Hidden Markov Model, Neural Networks Third Party Features etc. These are basically machine-learning methods.

Deep Natural Language Analysis Methods

In this category, all approaches involve complex natural language analysis techniques. None of these approaches solves the problem using machine learning, but rather uses a set of heuristics to create document extracts.

2.2 Multi-Document Summarization

Extraction of a single summary from multiple documents has gained interest since mid 1990s, most applications being in the domain of news articles. Several Web based news clustering systems were inspired by research on multi-document summarization, for example Google News, Yahoo News etc. This departs from single-document summarization since the problem involves multiple sources of information that overlap and supplement each other, being contradictory at occasions. Therefore, the key tasks are not only identifying and coping with redundancy across documents, but also recognizing novelty and ensuring that the final summary is both coherent and complete.

Abstraction and Information Fusion:

SUMMONS (Radev and MCKeown, 1998) is the first historical example of a multi-document summarization system. It takes multiple documents about a single event of narrow domain from various sources and produces a brief summary containing information about the event. Rather than working with raw text, SUMMONS reads a database previously built by a template-based message understanding system. The architecture of SUMMONS consists of two major components: a content planner that selects the information to include in the summary through combination of the input templates, and a lin-

guistic generator that selects the right words to express the information in grammatical and coherent text.

Graph Spreading Activation:

Mani and Bloedorn, 1997 describe an information extraction framework for summarization, a graph-based method to find similarities and dissimilarities in pairs of documents. Although no textual summary is generated, the summary content is represented via entities (concepts) and relations that are displayed respectively as nodes and edges of a graph. Rather than extracting sentences, they detect salient regions of the graph via a spreading activation technique. A document is represented as a graph as follows:

Each node represents the occurrence of a single word (i.e., one word together with its position in the text). Each node can have several kinds of links: adjacency links (ADJ) to adjacent words in the text, SAME links to other occurrences of the same word, and ALPHA links encoding semantic relationships captured through WordNet. Besides these, PHRASE links bind together sequences of adjacent nodes, which belong to the same phrase and NAME, and COREF link stands for co-referential name occurrences.

Centroid-Based Summarization

Generally this type of approaches do not use a language generation module(Zhang and Li, 2009). All documents are modeled as bags-of-words. The first stage consists of topic detection, whose goal is to group together news articles that describe the same event. To accomplish this task, an agglomerative clustering algorithm is used that operates over the TF-IDF vector representations of the documents. The second stage uses the centroids to identify sentences in each cluster that are central to the topic of the entire cluster. The system is easily scalable and domain-independent.

3 Our Approach

We have developed a text summarizing method for Assamese, based on the use of a WordNet and a stop word list. Since no prior Assamese WordNet exists, we have to build the required WordNet for our experiment. To populate the Assamese WordNet database (Hussain et. al, 2011) (as there is no publicly available WordNet database for Assamese) we uses the following sources of data -

- Online Dictionary
- Chandrakanta Abhidhan

We also create a stop word list of 168 Assamese words. Since no Assamese WordNet is available on the internet, the WordNet database was very small; therefore, we added to it all the words of our test document by ourselves.

3.1 Preprocessing

In the file, *Put every sentence in a new line*. It will help to retrieve each sentence easily and it can be solved by breaking each sentence on some special character. *All the words should be in their root form*. This is needed because the WordNet contains only the root words of a language. To solve this problem we need a stemmer. Designing a stemmer is difficult because it involves lots of morphological analysis. In our experiment, we manually performed this job. If a word is not available in the WordNet due to not being in root form, we can adopt the following idea for finding similarity between such two words. If the word $W_1=a_1a_2a_3\dots a_n$ and $W_2=b_1b_2b_3\dots b_m$ and not present in the WordNet then the similarity between W_1 and W_2 is

$$\frac{2 * |\text{length of matching character sequence of } W_1 \text{ and } W_2|}{|\text{length of } W_1| + |\text{length of } W_2|}$$

For example

$$W_1 = \text{মৰম} (\text{Maram : love})$$

$$W_2 = \text{মৰমৰ} (\text{Maramar : of love})$$

$$\text{Similarity} = (2 * 3) / (3 + 4) = 0.85$$

In (Zhang and Li, 2009) the author propose a sentence similarity computing method based on the three features of the sentences, the *word form feature*, the *word order feature* and the *semantic feature*, using weights to describe the contribution of each feature of the sentence. Since our work is on Assamese language and it is *free word order language* we do not need to consider the *word order feature*. In the next section of this report, we discuss the similarity measures that we used.

To calculate the semantic similarity we used the Assamese WordNet (Hussain et. al, 2011). To find the similarity of two words we first arrange the WordNet entry of these particular words tree

wise. The tree structure is created according to the relation between the words. After that, we count the number of edges N between both the words. If there is no relationship between the words then N become very large (size of the WordNet). In that case, similarity will be approximately zero. But in case of synonym words there are no any edges between the words. But in that case similarity should be one. Therefore after deriving N we will calculate the similarity as follows.

$\text{WSS}(w_1, w_2) = 1/(N+1)$ i.e. semantic similarity will proportional to $1/N$.

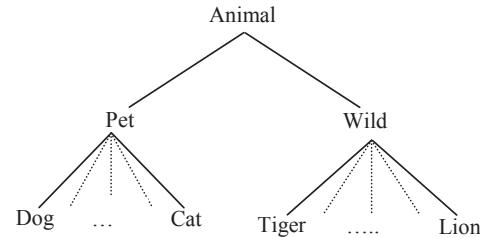


Figure 1: A fragment of WordNet

Similarity measure between sentences:

Definition 1: Word Form Similarity.

The word form similarity is mainly used to describe the form similarity between two sentences. It is the number of same words in two sentences measures it. First, we get rid of the stop words. If S_1 and S_2 are two sentences, the word form similarity is calculated by the formula (Zhang and Li, 2009).

$$\text{Sim1}(S_1, S_2) = \frac{2 * (\text{SameWord}(S_1, S_2))}{(\text{Len}(S_1) + \text{Len}(S_2))} \quad \dots \dots \dots \quad (1)$$

Definition 2: Word Semantic Similarity.

The word semantic similarity is mainly used to describe the semantic similarity between two sentences. Here the word semantic similarity computing is based on the Assamese WordNet. Based on semantic similarity among words, (Zhang and Li, 2009) Word-Sentence Similarity (WSSim) is defined to be the maximum similarity between the word w and words within the sentence S . WSSim(w, S) is defined with the following formula

$$WSSim(w, S) = \max\{Sim(w, W_i) / W_i \in S, \text{ where } w \text{ and } W_i \text{ are words}\} \quad \dots \dots \dots (2)$$

Here the $Sim(w, W_i)$ is the word similarity between w and W_i . With $WSSim(w, S)$, the sentence similarity is defined as follows:

$$Sim2(S_1, S_2) = \frac{\sum_{i=1}^{|S_1|} WSSim(w_i, S_2) + \sum_{j=1}^{|S_2|} WSSim(w_j, S_1)}{|S_1| + |S_2|} \quad \dots \dots \dots (3)$$

In (3) S_1, S_2 are sentences; $|S|$ is the number of words in the sentence S .

Definition 3: Sentence Similarity.

The sentence similarity is usually described as a number between zero and one, zero stands for non-similar, and one stands for totally similar. The larger the number is, the more the sentences similar. The sentence similarity between S_1 and S_2 is defined as follows (Zhang and Li, 2009):

$$Sim(S_1, S_2) = \lambda_1 * Sim1(S_1, S_2) + \lambda_2 * Sim2(S_1, S_2) \quad \dots \dots \dots (4)$$

In (4) λ_1 and λ_2 are constants, and satisfy the equation $\lambda_1 + \lambda_2 = 1$. λ_1 and λ_2 defines the contribution of the semantic similarity and word form similarity between S_1 and S_2 . In our implementation we assumed $\lambda_1 = \lambda_2 = 0.5$.

The function $Sim()$ (Equation (4)) is used as the final measure of sentence similarity. Since we focus mainly on sentence clustering, we need a proper similarity function, which will measure the similarity between two sentences not only in terms of word level but also in semantic level. In this work during sentence clustering, we use this function as the similarity function. Next, we need to find the number of clusters.

3.2 Estimating the number of clusters

Determination of the optimal number of sentence clusters in a text document is a difficult issue and depends on the compression ratio of summary and chosen similarity measure, as well as on the document topics. For clustering of sentences, author used a strategy to determine the optimal number of clusters (the number of topics in a document) based on the distribution (Das and Martins, 2007) of words in the sentences:

$$k = n \frac{|D|}{\sum_{i=1}^n |S_i|} = n \frac{\left| \bigcup_{i=1}^n S_i \right|}{\sum_{i=1}^n |S_i|} \quad \dots \dots \dots (5)$$

Where $|D|$ is the number of terms in the document D , $|S_i|$ is the number of terms in the sentence S_i , n is the number of sentences in document D . Here terms refers to all those words which are not in stop word list. From the above formula we can see that if the number of terms which are common to some sentences are increased then the number of clusters will be reduced, i.e. common terms in multiple sentences means the domain of the document is small. Such documents contain less number of topic. Here we analyze the property of this estimation by two extreme cases.

Case 1:

The document is constituted of n sentences, which have the same set of terms. i.e. all the sentences are constituted of the same words. Therefore, the set of terms of the document coincides with the set of terms of each sentence i.e. $D = (t_1, t_2, \dots, t_m) = S_i = S$. From (5) it follows that

$$k = n \frac{\left| \bigcup_{i=1}^n S_i \right|}{\sum_{i=1}^n |S_i|} = n \frac{\left| \bigcup_{i=1}^n S \right|}{\sum_{i=1}^n |S|} = n \frac{|S|}{\sum_{i=1}^n |S|} = 1$$

Case 2:

The document consists of n sentence which do not have any term in common, that is, $S_i \cap S_j = \emptyset$ for $i \neq j$. This means that each term belonging to D belongs only to one of the sentences S_i . i.e.

$$|D| = \left| \bigcup_{i=1}^n S_i \right| = \sum_{i=1}^n |S_i|$$

Therefore from (5) it follows that $k=n$. In both the extreme cases are depicted correctly. We assume that it will also work at any intermediate state. Therefore, we use the formula to find out the number of topics or the number of clusters of sentences in the document.

3.3 Summary Generation

After calculating the number of clusters, we use the K-means algorithm to cluster the sentences of the document. After clustering the sentences of the input document, the following few steps are there to find the final summary.

Step1: Extract the central sentences of each cluster.

Step2: Find similarity between headline (title) and those sentences, which are not included in the Step 1.

Step3: Add those sentences that are highly similar to the headline.

Step4: Sort them according to occurrence in the original input document.

Step5: Put the sorted sentences into the output document.

Based on the result of clustering, suppose the sentences clusters are $D = \{C_1, C_2, \dots, C_k\}$. First, determine the central sentence μ_i of each cluster based on the accumulative similarity between the sentence S_i and other sentences, and then calculate the similarity between the sentence S_i and the central sentence μ_i . Assume that the similarity of central sentence μ_i as 1, sort the sentences based on their similarity weights, and choose the high weight sentences as the topic sentences. After finding out the topic sentences, we add some more sentences to the summary based on the high similarity with the headline. In the summary generation process, there are mainly two steps. In the first stage, we select the cluster sentence, which is used as pruning of duplicate data. And in the second stage i.e. in step 2 and 3 we add those sentence which are representing the main topic of the document. Our entire method contains three phases as follows-

Phase1: Stem the words to obtain their root forms.

(Performed Manually in our experiment)

Phase2: Create a table of similarity between every sentence. (Auto)

Phase3: Cluster the sentences. (Auto)

Phase4: Summary Generation (Auto)

4 Experiments and Results

We conducted experiments to evaluate the performance of the automatic text summarization system based on sentences clustering. Automatic text summarization systems and evaluations of

summary is not a straight-forward process. For evaluate the results we use F-measure which is widely used in Information Retrieval. Due to lack of large WordNet database, we restrict our experiment to a few small documents as input to find the summary. Table 1 gives the obtained result of 10 media documents with different domain.

Document	P	R	F
1	0.73	0.84	0.78
2	0.75	0.75	0.74
3	0.82	0.93	0.87
4	0.79	0.73	0.75
5	0.81	0.86	0.83
6	0.85	0.77	0.8
7	0.82	0.77	0.79
8	0.75	0.8	0.77
9	0.87	0.9	0.88
10	0.72	0.61	0.66

Table 1: Obtained Result

5 Conclusion

We have presented the approach to automatic text summarization based on the sentences clustering and extraction. The main contribution of this report is that it proposed and implemented a sentence similarity computing method based on the semantic features of the sentences, based on analyzing the word form Assamese WordNet (Hussain et. al, 2011). We find that the approach produces good result and can be considered for further improvement.

References

- D. Das and A. Martins, “A Survey on Automatic Text Summarization”. Literature Survey for the Language and Statistics” 2007
- C.-Y. Lin, and E. Hovy, “Identifying topics by position”, In Proceedings of the Fifth conference on Applied natural language processing, PP: 283-290, 1997
- D. R. Radev, and K. McKeown, “Generating natural language summaries from multiple online sources”. Computational Linguistics, 24(3), PP: 469-500, 1998

- I. Mani, and E. Bloedorn, "Multi-document summarization by graph search and matching. In AAAI/IAAI, PP 622-628, 1997
- P. Zhang and C. Li., Automatic Text Summarization Based on Sentences Clustering and Extraction. In Proceedings of IEEE International Conference on Computer Science and Information Technology, PP 167–170, 2009
- I. Hussain, N. Saharia, U. Sharma "Development of Assamese WordNet" Machine Intelligence:Recent Advances. Narosa Publishing House, Editors. B. Nath, U. Sharma and D.K. Bhattacharyya, ISBN-978-81-8487-140-1, 2011

Online Multilingual Dictionary of North-East India, <http://www.xobdo.net>

Chandrakanta Abhidhana (Online),
[http://dsal.uchicago.edu / dictionaries/candrakanta/](http://dsal.uchicago.edu/dictionaries/candrakanta/)

Assamese Vocabulary and Assamese Wordnet Building: An Analysis

Shikhar kr. Sarma

Dept. of Computer and IT
Gauhati University
Guwahati, India 781014
sks001@gmail.com

Mayashree Mahanta

Dept. of Computer Science
Gauhati University
Guwahati, India 781014
mayashreemahanat@gmail.com

Utpal Saikia

Dept. of Computer Science
Gauhati University
Guwahati, India 781014
utpal.sk@gmail.com

Himadri Bharali

Dept. of Computer Science
Gauhati University
Guwahati, India 781014
himadri0001@gmail.com

Abstract

The present research paper is an attempt to provide an idea of vocabulary system in Assamese language. It also tries to highlight the conditions under which Modern Assamese language was separated from Indo-Aryan Language in 10th century A.D. It also deals with the enrichment of the vocabulary of Assamese language through the ages. Wordnet is contributed much in the development of Assamese vocabulary. In fact, this paper will also point out how Assamese vocabulary helps in building the Wordnet in Assamese with suitable examples.

Keywords: Assamese Wordnet, Vocabulary.

1.0. Introduction

Vocabulary is the backbone of any language. It expands the horizon of any language and builds a close relation among various languages. Thus, Assamese language also has gained a large amount of words from various languages by different processes in course of time. Assamese language was established as an independent language since 10th century A.D. and in 14th century A.D., the written form of the language could be found. The development of the language and literature in Assamese till today has greatly contributed to the growth and enrichment of the vocabulary of the language.

Wordnet is prepared to establish a bond of understanding among the various languages of the

world. In this respect, Assamese also plays a major role in building Wordnet of Assamese language. As a result, Assamese Wordnet is being prepared on the basis of Hindi Wordnet.

1.1. A discussion of Assamese vocabulary and Wordnet (Assamese Wordnet) building is given in following:

1.2. The scope of Assamese vocabulary is very vast. It consists of words of Sanskrit origin, Non-Aryan words, dialect oriented words. Besides Assamese socio-cultural influences are also perceived in the vocabulary of the language. It is to be noted here that in this paper we would basically deal with the standard Assamese language forms. Assamese still lacks a common vocabulary dictionary in the language. Moreover, no dictionary was found in the early and the middle ages. The selected modern dictionaries are – ‘A Dictionary in Assamese and English by Miles Bronson’ (1867); ‘Hemkosh’ (1900), by Hemchandra Barua and later it is compiled by Debananda Barua (the 14th edition) which included 1,54,428 words; ‘Chandrakanta Abhidhan’ (2004, 3rd edition) ‘Adhunik Asomiya Sabdakosh’ (2007, 9th edition), ‘Asamiya Jatiya Abhidhan’ (2010) and many other vocabulary dictionaries are available in Assamese language. No common standard vocabulary dictionary has been made till today. Many critics have prepared vocabulary lists in their own way. Earlier philologists like Kaliram Medhi and Banikanta kakati had classified vocabulary list in their own style.

2.0. Kaliram Medhi in ‘Asomiya Byakaran aru Bhasatattva’[3] has provided a classification Assamese vocabulary such as ‘tatsama’, ‘tatbhava’ and ‘desaja’. But his ‘Desaja’ words are shown as loan words in which maximum words are Per-

so-Arabic words. Therefore, his vocabulary classification cannot be taken as valid. On the other hand, though Banikanta Kakati's classification[1] of Assamese vocabulary covers almost all the aspects, yet his classification also cannot be regarded as valid one.

2.1. It is interesting to note here that there are a large amount of loan words in Assamese language. In day-to-day life these loan words have been used extremely to express feelings, ideas etc. Moreover, it is seen that Perso-Arabic words have been used in Assamese language. these words occupy a significant status in Assamese language.

3.0. Assamese vocabulary

Assamese vocabulary can be divided into the following heads [5]

1. Aryan or words of Sanskrit origin
 - a.Tatsama
 - b.Semi-tatsama
 - c.Tadbhava
2. Non-Aryan words
 - a.Austro-Asiatic
 - b.Tibeto-Burmese
 - c.Tai-Ahom
 - d.Dravidian
3. Loan words
 - a.Words coming from N.I.A. Languages
 - b.Foreign Words
 - i.Persian
 - ii.Arabic
 - iii.Portuguese
 - iv.English
 - c.Loa translations
 - i.Translated words
 - ii.Terminology
4. Unclassified words
 - a. Hybrid
 - b. Onomatopoetic
 - c. Compound

3.0.1. Sanskrit origin

The development of the Assamese language is very much indebted to the Sanskrit. Nevertheless, other languages also play an active role in contributing to the development of vocabulary in the language, but Assamese can be characterised as a Indo-Aryan language only due to Sanskrit.

3.0.1.1. Tatsama words

Though Assamese originated from Magadhi Prakrit, many tatsama words have been used in the language. No tatsama words are seemed to be found in Assamese. However, when a language takes words from another language, there occurs some changes in the pronunciation in the borrowed words. These kinds of changes are very much frequent in Assamese. In Sanskrit, there were 11 vowel sounds and 8 in Assamese. Like Sanskrit, Assamese does not possess Dental, palatal and labial sounds, but instead it has dental sound in Assamese. There is no difference between ‘સ’/s/ and ‘છ’/ts/ or ‘જ’/z/ and ‘ઝ’/dz/. These can be termed as dental sounds. The pronunciation of the sounds like શ, ષ, સ /x/ can be found in more changed forms. Besides, as many of Sanskrit sounds are not available in Assamese, so we cannot find the pronunciation of Sanskrit words in Assamese language. Some examples of Assamese *tatsama* words:

/gobyo/	(related to cow)
/ xoisjo/	(grain)
/vntorzami/	(god) etc.

In fact, there are certain grounds in support of the availability of tatsama words in the Assamese language. These are mentioned below[5]:

- 1 We find all the Sanskrit letters in Assamese spelling system as it is written on the basis of Sanskrit spelling system, though in speaking these are not used in pronunciation.
- 2 These words are determined according to the grammatical rules of sandhi, samasa etc. of Sanskrit.
- 3 In these words, it is seemed to follow the rules of natva-vidhi and satva-vidhi.
- 4 In case of tatsama words, it is found to add -tara, -tama and istha in degree of comparison. But in case of tatbhava words such suffixes are not used.
- 5 In the use of tatsama words in Assamese, some conventional norms are employed. In compounding, both the words must be either tatsama or tadbhava and never a mixture of both.
- 6 Moreover, tatsama words are used depending on the context of the sentence structure.

3.0.1.2. Semi-tatsama words

Semi-tatsama or Ardha-tatsama words mean half of the word is equivalent to that of Sanskrit. Unlike tadbhava words, these words are new in the Assamese language and retain the traces of the original form. Besides, semi-tatsama words come through the Prakrit stage. There are some processes which help in the changes of these words like Anaptyxis, epenthesis and metathesis. In fact, the number of semi-tatsama words is very less in comparison to tatsama and tadbhava words. There are some examples of Assamese semi-tatsama words given below:

/kɔrɔm/ (work)
 /dɔɔrɔm/ (duty)
 /duwar/ (door) etc.

3.0.1.3.Tadbhava words

Tadbhava words are those words which were handed down through the M.I.A. stage. In the M.I.A. period, the O.I.A. words went through a significant change and again were changed in the N.I.A. period. There are some examples of tadbhava words given below –

/boai/ < / bɔrata/	(brother)
/loha/ < /louhv/	(iron)
/misa/ < /mitvya/	(false) etc.

3.1.

Assamese has been used as a link language in North-east India. Therefore, it is found that Assamese contains words from other neighboring language families. For this reason, Assamese can be regarded as a separate language from other Indo-Aryan language. The contributions of these Non-Aryan languages into Assamese make it able to keep the greater Assam as one family. This is observed from the following examples.

3.1.1: Non-Aryan words

- (a) Austro- asiatic: / nodo-ka/(fleshy), /sikora/(a tick), /selauri/(eye-brow) etc.
- (b) Tibeto-Burman: /gospok/(act of trampling) , /gɔba-mar/(hug),

/ tɔplamuri mar/ (to slap on the head with the palm of the hand)etc.

- (c) Tai-Ahom: / maihan/(a dish), /karɔŋ/(palace),/ burɔnji/ (history)etc.
- (d) Dravadian:/ amoi/(a motherly woman), /urukp/(to leak), /hɔp/(to draw into the mouth or smoke), etc.

3.1.2: Loan words:

In Assamese, we can see many loan or borrowed words from other Indo-Aryan, foreign languages. There are many reasons behind the coming of the loan words in Assamese. It may be due to social contact or due to the study of literature etc. In fact, Assamese has been enriched with a large numbers of non-native words through the ages and they came to be used native-like words.

3.1.2.1.Words coming from other N.I.A. languages —

- (i) Bangla : / rɔsvgulla/(one kind of sweet),/ xɔndox/(one kind of sweet), /kʌlʌŋkari/(scandal),/ bɔvjal/(adulterated) etc.
- (ii) Hindi : /kagɔjwala/(a hawker), /koadi/(an woman cloth of hand yarn), /koali/(empty), /dɔkait/(dacoit) etc.
- (iii) Marathi : /gaoburha/(a village head-man), /bɔrɔŋni/(contribution),/ bɔŋji/(niddle), /taŋɔrɔn/(edition) etc.
- (iv) Gujarati : /hɔrtal/ (protest)etc.

3.1.2.2. Foreign words

- (i) Persian: / aina/(mirror), /adalɔt/(court), /dokan/(shop), /dɔrji/(tailor),/ adar/(honour)
- (ii) Arabic: /dowat/(ink pot),/ hajira/(presence),/ hakim/(judge),/tabij/(charm)
- (iii) Portuguese: / saku/(knife), /daroga/(constable), /bɔŋgam/(queen),/ anarɔx/(pineapple),
- (iv) English: /pluto/, /meter/, / ameba/, / bus/, /tragedy/, / comedy/
- (v) Other foreign words-
 - French:/ boutique/, /cafe/
 - Dutch:/ hارتان/, /rahittan/, /iskapan/, /turup/ etc.
 - Spanish: /cigarette/

3.1.2.3. Loan translated

- (i) Translated words:
 /durdurxɔn/(television),/konduwa-gps/(tear-

- gas),/ batori kakot/ (news paper) / kola - bojar/ (black market)
- (ii) Terminology: Various types of terminological words are used in Assamese language whenever necessary.

3.2. Unclassified words

There are certain words in Assamese language which are labeled as unclassified. These are:

- (i) Hybrid:/ pññit-giri/(fancying oneself to be learned), /tikot-gvñr/(ticket house),/ basastan/(bus station),/ man-ijjot/ etc.
- (ii) Onomatopoetic: / pouspousñni/ (whispering), / kouji-magi/(act of begging),/ dabi-dvñmki/ etc.
- (iii) Compound: / atou-karh/(to kneel down), /hat-bojar/(market), / rñndba-bñrha/(cooking) etc.

4.0. Use of Assamese vocabularies in Assamese Wordnet

Wordnet is basically a synonymous lexical database. Vocabulary plays a main role in building Wordnet. As Assamese language possesses a huge amount of vocabulary, it becomes easy to build Wordnet in the language. The task of Assamese Wordnet building is almost ready to provide us with all the lexical words. Yet there are still many words in the language those need to be entered.

Assamese Wordnet is built on the basis of Hindi Wordnet. Here, words are shown according to the sense of the given context or sentence and accordingly, we can derive different meanings from them. For example, the assamese word ‘*bhaal*’ has different meaning according to its sense in the context.

Bhaal (Adj) - *jijan bhaal gunar adhikaari*
-jito sakalo phaalar paraa bhaal abasthaat thaake
-jito bhaangi-singi baa beyaa hoi juwaa nai
-jito shubha baa bhaal hay
-dekhoot shuwani
- su-swaad jukta
-gun samriddha
-jaar parinati shubha

Apart from sense, we also deal with the morphological structure of the words. The grammatical category of a given word is changed by using

morphological processes. Generally, affixes are used to show the different grammatical category of the words in different context. The following examples represent it clearly:

dakshya (n)- dakshyajane kaamto kariba
dakshya (adj)- ei kaantor babe tekhet dakshya
dakshya(adv)- tekhete kaamto dakshyataare sampanna karile.

5.0. Conclusion

This paper highlights the Assamese vocabulary system and also the use of vocabulary in building Assamese wordnet. Though Assamese wordnet tries to cover all the Assamese vocabulary, yet it cannot able to represent all the forms of the vocabulary of Assamese language.

Acknowledgement: The authors acknowledges the TDIL programme of Department of Information Technology, Ministry of Communication and IT of Government of India for funding the Wordnet Project on Assamese Wordnet Development currently being executed at Gauhati University.

References

- Banikanta Kakati. 2008. Assamese : Its Formation and Development, Lawyers Book Stall, Guwahati , Assam.
- Golock C. Goswami. 1983. Structure of Assamese, Gauhati University , Assam
- Kaliram Medhi 1999. . Asamiya Byakaran Aru Bhassattva, Lawyers Book Stall, Guwahati, Assam.
- Lilabati bora Saikia. 2006. Asamiya Bhasar Rupatattva, Banalata, Guwahati, Assam.
- Ramesh Pathak. 2004. Studies in Assamese Vocabulary, Anita Pathak, Guwahati, Assam.
- Shikhar Kr. Sarma, Utpal Saikia & Mayashree Manta. 2010. Kinship Terms in Assamese Language, Department of Computer Science, Gauhati University, Assam 3rd Wordnet Workshop , IIT Kharagpur, India.
- Shikhar Kr. Sarma, Moromi Gogoi, Rakesh Medhi & Utpal Saikia.2010. Foundation and Structure of Developing an AssameseWordnet, Department of Computer Science Gauhati University, Proceedings of the 5th Global Wordnet Conference, Narosa Publishing House.

Encoding Commonsense Lexical Knowledge into WordNet

Gianluca E. Lebani^{†‡}

[†]Center for Mind/Brain Sciences
University of Trento, Italy
gianluca.lebani@unitn.it

Emanuele Pianta^{†‡}

[†]HLT Group
Fondazione Bruno Kessler, Italy
pianta@fbk.eu

Abstract

In this paper, we propose an extension of the WordNet conceptual model, with the final purpose of encoding the common sense lexical knowledge associated to words used in everyday life. The extended model has been defined starting from the short descriptions generated by naïve speakers in relation to target concepts (i.e. feature norms). Even if this proposal has been developed primarily for therapeutic purposes, it can be seen as a generalization of the original WordNet model that takes into account a much wider and systematic set of semantic relations. The extended model is also an enhancement of the psycholinguistic vocation of the WordNet model. A featural representation of concepts is nowadays assumed by most models of the human semantic memory. For testing our proposal, we conducted a feature elicitation experiment and collected descriptions of 50 concepts from 60 participants. Problematic issues related to the encoding of this information into WordNet are discussed and preliminary results are presented.

1 Introduction

WordNet (WN: Fellbaum, 1998) is the largest and most systematic lexical database in electronic format available nowadays. Nevertheless, it is hard to maintain that such a successful and widely used resource contains a complete (or near-to-complete) representation of the information that is encoded in the mental lexicon of English speakers. The lack of completeness is not only referred to the coverage of lexical unites, or to the population of the already defined lexical and semantic relations (see for instance the sparse instantiation of the meronym relation), but also to structural aspects, such as: the number and type of the encoded relations; the encoding of the strength (or any similar quantitative notion) of relations, in order to represent, for instance, pro-

totypicality effects; the encoding of quantifiers and logical operators, as an important aspect of the knowledge associated to concepts; the encoding of syntagmatic information, e.g. collocations and selectional preferences or restrictions.

In order to overcome such limitations, in the last twenty years, many extensions of the WN conceptual model have been proposed by the creators of the original Princeton WordNet (PWN) and by other scholars in projects such as EuroWordNet (EWN: Alongs et al, 1998), MultiWordNet (MWN: Pianta et al, 2002), WordNet Domains (Bentivogli et al, 2004) and BalkaNet (Tufiş, 2004). However, none of such proposals has tried to define, on the basis of psycholinguistic evidence, a close set of semantic relations that are expected to be able to represent all (or most) of the meaning aspects conveyed by a concept. In this paper we make such an attempt, starting from the requirements of a very specific application scenario, in which an electronic lexical database is used to support speech therapists in their daily work with aphasic patients.

2 Background and Motivation

Anomia is a common symptom associated with aphasia. Most patients affected by an acquired linguistic disorder due to a brain damage experience some difficulty in retrieving or producing words. In this context, computers can be helpful in many ways: from assisting the therapist in the rehabilitation, to helping the patients in his/her everyday life (Petheram, 2004). Given the great variability of forms and severity in which anomia can manifest itself, a requirement that any assistive tool has to meet is to be flexible enough to fit the needs of different classes of patients.

STaRS.sys (Semantic Task Rehabilitation Support system) is the outcome of a joint effort between Fondazione Bruno Kessler and the CIMEC Center for Neuropsychological Rehabilitation (CeRIN). The aim of this project is the crea-

tion of a tool for supporting the therapist in the preparation of rehabilitative tasks for Italian-speaking patients affected by anomia.

Typically, the information exploited in semantic rehabilitation tasks can be represented as concept-description pairs like *<chair> has four legs*, *<airplane> flies*¹. Notably, this is the same kind of information that is collected by scholars who study the characteristics of conceptual knowledge by running *feature generation* experiments, that is by asking speakers to describe concepts (cfr. Murphy, 2002). We've argued elsewhere (Lebani and Pianta, 2010b) that the WordNet conceptual model fits well the STaRS.sys requirements, so that we chose to build the STaRS.sys semantic knowledge starting from the Italian MWN lexicon (iMWN).

We believe that only a lexicon organized on the basis of psycholinguistic evidence can be flexible enough to meet the STaRS.sys requirements. As a matter of fact, many psycholinguistic assumptions lay at the basis of the WN model (e.g. Miller, 1998), and its psychological validity has been tested explicitly or implicitly by several scholars (e.g. Fellbaum, 1998b; Izquierdo et al, 2007; Barbu and Poesio, 2008). However, just few of the many WN extensions proposed in the last two decades seem to be based on psycholinguistic hypotheses and methodologies.

An outstanding exception to this trend is the evocation relation by Boyd-Graber et al (2006), who proposed the introduction of weighted, oriented arcs between pairs of synsets, e.g. from {car} to {road}, representing how much a concept evokes the other. The relation has been populated by collecting judgments from speakers (Boyd-Graber et al, 2006; Nikolova et al, 2011).

There are many similarities between our work and that by Boyd-Graber and colleagues. In both proposals, WN is enriched with speaker generated semantic information, and the encoding of this information requires an extension of the WN model. Also, both proposals are exploited for assistive purposes. The resource by Boyd-Graber and colleagues has been adopted as the semantic knowledge base behind the tool ViVa (Nikolova et al, 2009), a visual vocabulary designed for aiding anomic patients in their everyday life. In a similar way, STaRS.sys will be part of a comput-

er aided therapy tool designed for supporting therapists in their daily work with patients. In spite of the commonalities between the two projects, we observe that a generic evocation relation seems not to meet all the requirements of speech therapists, which need instead a more fine-grained classification of semantic relations. For the STaRS.sys purposes, we need to encode structured lexical information that is more similar to what can be obtained by exploiting a feature generation paradigm, than to what can be obtained through free associations.

Due to the great variability of impairment shown by anomic patients and to the lack of resources, the preparation of a therapeutic task for anomia rehabilitation is a manual work on behalf of the therapist. STaRS.sys is a system thought for being helpful in this preparatory phase by helping the therapist to (1) retrieve concepts, (2) retrieve information associated to concepts and (3) compare concepts. In the knowledge base underlying this system, the following kinds of information have to be available for every concept: its position in a conceptual taxonomy; a set of featural descriptions (FDs) classified according to the types of knowledge conveyed; a value of prototypicality and of word frequency.

As argued in Lebani and Pianta (2010b), the WN conceptual model fits well our needs, because of its cognitive plausibility, for its ease of use and because it is based on a fully specified is-a hierarchy. Moreover, it is powerful enough, with some modifications, to represent the information contained in featural descriptions. FDs like *<cup> is used for drinking* can be represented in WN as a relation (say *is Used for*) holding between the described (or "source") synset {cup} and the most prominent synset of the description, i.e. the ("target") synset {drink}.

A similar assumption has been used by Barbu and Poesio (2008), who analyze the overlap between the semantic information encoded in PWN and in the collections by McRae et al (2005) and Garrard et al (2001). In their analysis, the authors, who also considered information contained in glosses, estimated that the overlap between PWN and existing norms collections can vary between 22 and 40% (depending on the collection and on the method used to calculate the overlap). The same analysis showed that the WN coverage with regards to FDs is highly skewed (e.g. categorical information is highly present, whereas functional information is missing).

To overcome some of the limitations of the current WN model, Lebani and Pianta (2010b)

¹ Concepts and features are printed in *courier new* font. When reporting a concept-feature pair, the concept is further enclosed by *<angled brackets>*. WordNet synsets are enclosed by *{curly brackets}*. Feature types, relations and concept categories are reported in *italics times roman*.

proposed to add a set of 25 semantic relations in a dedicated version of iMWN called StarsMultiWordNet (sMWN), with the final objective of finding a complete set of intuitive and cognitively plausible relations representing lexical meaning. This extension has been built by combining experimental evidence from existing feature norms with theoretical proposals developed in lexicography, linguistics and cognitive psychology (for details, see Lebani and Pianta, 2010a).

This paper presents the results of a pilot study aiming at populating the extended set of WN relations by collecting FDs from subjects in a controlled setting, and encoding them into sMWN. Section 3 will present available feature norms collections; Section 4 will illustrate the results of the collection experiment and Section 5 will comment on the issues faced when actually mapping FDs into WN relations.

3 Available feature norms collections

Since the early times of the cognitive psychology enterprise, the feature norm paradigm has been widely employed in the investigation of the human's conceptual representation and computation (cfr. Murphy, 2002). Despite this wide use, to date there are few freely available collections (Garrard et al, 2001; McRae et al, 2005; Vinson and Vigliocco, 2008; De Deyne et al, 2008; Kremer and Baroni, 2011). These resources are strongly influenced by the goals and theoretical framework of the connected studies, so that they differ substantially on the quantity and kind of described concepts, on the procedure adopted for collecting and processing features and on the classification adopted for classifying them.

In the canonical paradigm, speakers are simply asked to describe a concept. On the one side, this approach has shown his utility for investigating which concepts and/or properties are easier to recall. On the other side, however, it produces a very sparse population of the various components of lexical meaning. As an example, consider that 75.44% of the descriptions of the McRae dataset belongs to just 7 types out of 27. Many factors may contribute to this sparseness, among which the organization of the human semantic memory itself. It is also probable, however, that part of this disproportion is due to the methodology exploited for eliciting and normalizing descriptions. Because of the sparseness of property types, it turned out that none of the available collections can be efficiently exploited for our purposes, as we need to collect FDs that are as va-

ried as possible. We coped with this issue by adopting a question answering paradigm for the elicitation experiment, as described in Section 4.

Another problematic issue in existing collections concerns the normalization of raw descriptions. Even if this practice is claimed to be as much conservative as possible, the ways in which it is usually carried out leads, from our point of view, to a loss of knowledge. Furthermore, our feeling is that too much is left to the interpretation of the persons in charge of the normalization. As an example, in the Kremer norms, the description of the pair *<garage>* can be used as a utility room is paraphrased as used for storing. However in this way we miss the information that *garage* and *utility room* are similar concepts, encoded by the coordination relation in our relation scheme we will show in Section 5 how iMWN can be used to alleviate such problems.

4 A new norms collection

Given the limitation of existing norms collections, we decided to conduct an elicitation experiment adopting the stimulus set by Kremer and Baroni (2011) and a comparable number of participants, with a slightly different methodology. This allows for the comparison of our dataset with the only freely available norms in Italian.

4.1 Experimental Setup

Participants: 60 Italian speakers participated in the the experiment. Their age ranged from 19 to 55 years (mean: 28.9, s.d. 9.27). All subjects were recruited in the university environment.

Materials: The stimulus set was composed by 50 concepts belonging to the following 10 categories: *bird*, *body part*, *building*, *clothing*, *fruit*, *furniture*, *implement*, *mammal*, *vegetable* and *vehicle*. Kremer and Baroni (2011) selected these same 50 concepts for the reasonable unambiguousness of their lexical realizations.

Procedure: The descriptions have been collected through an on-line experiment. 12 groups of 5 tasks were prepared, each task composed of 10 randomly ordered concepts, one for each category. In this way, every concept has been described by 12 subjects, and no participants received a questionnaire that was previously assigned to another participant.

The semantics of each relation has been paraphrased as a question of the form: “*what are the portions of a [concept]?*” (for the *hasPortion* relation). This allowed us to populate as much as

possible all feature types, and to reduce need for interpretation in the normalization process.

Every subject has been presented a concept per web page, followed by a set of relevant questions. For each question, examples were available in the online documentation, accessible by clicking on the question text. Subjects were instructed not to report any biographic or technical knowledge, and they were allowed to leave a field empty if they didn't come up with any answer. Participants were trained on two example concepts (*cat*, *knife*) for which some suggestions were supplied in different ways (pre-filled fields, auto-completion).

4.2 Results

We collected 18,884 raw FDs, that is a mean of 377.68 descriptions (s.d. 60.71) per concept. Every subject, on average, produced 314.73 (s.d. 115.68) descriptions over 10 concepts and 31.47 descriptions per concept (s.d. 13.71).

In a pre-processing phase every FD has been analyzed as an instance of one the feature types proposed in Lebani and Pianta (2010b). In doing so, we exploited the fact that all FDs have been produced as an answer to a specific question that was formulated on the basis of one these feature types. The appropriateness of the descriptions was manually checked by one of the authors. This led to the deletion of 1,023 raw descriptions because they were conveying technical, autobiographical or patently wrong information. Given the remaining descriptions, in 2,247 cases we re-categorized the FD, and associated it to a feature type different from that implied by the subject. Summing up, a total of 3270 features (17.3% of the total) underwent some change in this phase.

Comparison with the Kremer norms: A preliminary quantitative evaluation of our dataset shows that we have collected 18,884 descriptions against 8,250 descriptions in Kremer dataset. Other meaningful comparisons concern the number of descriptions per subjects (314.73 vs. 123.48), the number of descriptions per concept (377.68 vs. 170.4) and the average of feature per concept produced by every subject (31.47 vs. 4.96). These data suggest that our strategy paid off, by providing a richer and more systematic set of feature descriptions for each concept.

5 Encoding descriptions into WN

The second step of our pilot study consisted in manually populating sMWN with the *normalized* version of the 1,785 raw descriptions collected

for the following five concepts: *seagull*, *finger*, *chair*, *corn* and *airplane*.

5.1 The encoding procedure

The manual encoding of the FDs content in sMWN is based on two main criteria. First, the annotator should have minimum space for interpreting the data. Second, the simplification of the informative content of a description should be used only as a “last resort” strategy.

Normalization: In works belonging to the feature generation paradigm, the collection of the descriptions is always followed by a normalization step, in which semantically equivalent FDs are merged. However, often a clear explanation of how equivalent descriptions are identified is missing. As an example, raw descriptions like *is a quadruped* and *has four legs* can be seen as exemplars of the same feature (e.g. *has four legs*) and merged (cfr. Vinson and Vigliocco, 2008). It is questionable, however, that these expressions convey the same information. A quadruped is “an animal that moves by using four legs”, and reducing its definition to “having four legs” is reductive.

In our approach equivalent descriptions are defined as descriptions sharing the same semantic relation and the same source and target synsets. Accordingly, then, we consider the two FDs *<wheel> is a component of a car* and *<wheel> is an auto part* equivalent because they can be both mapped into a meronymic relation linking {*wheel*} and {*car*, *auto*}.

Ambiguity: In a number of cases the FD contained an ambiguous word, so we need to choose an appropriate synset for it. We identified two variants of this situation.

If the concurrent synsets are in a hyponym relation, and the property is possessed by all the hyponyms of the more general synset, this is selected. As an example, the target concept of the FD *<coltello> è usato dal cuoco* (*<knife>* is used by the cook/chef) can be represented in sMWN as the Italian equivalent of either {*cook*} or {*chef*}, where the first is a hypernym of the second. In this situation, given that the property of “using a knife” is possessed by all hyponyms of {*cook*}, our choice falls on the more general synset.

Instead, when the property cannot be predicated of all the hyponyms of the more general synset, we opt for the more specific. Consider the pair *<ciliegia> cresce in giardino* (“*<cherry>* grows in gardens/grounds”).

The target concept, in this case, can be encoded with the Italian translations of both {grounds} and {garden}. However, since cherry trees to now usually grow in a {parvis} or in other hyponyms of {grounds} according to sMWN, we encoded this feature as a relation holding between {cherry} and {garden}.

In most cases the synsets corresponding to the ambiguous words are not one the hyponym of the other. As an example, given the FD <corn> can be found in a cellar, the target concept cellar can be encoded as either {basement, cellar} or {root_cellar, cellar}. Given that both synsets look plausible, we chose to double the concept-description pair in the database.

Loose Talk: Speakers are not dictionaries, so they may ignore some terms or they simply may not recall them in a certain moment. As a consequence, some raw phrases express concepts that could be expressed by an existing term, such as *is used by people who cook*.

In the standard feature generation paradigm, descriptions like these can be interpreted in many ways. They may even be re-phrased as features of a different kind, such as *is used for cooking*. In our approach, the rephrasing is guided by the synsets and glosses available in WN. In our case, we choose the synset {cook} given the gloss “*someone who cooks food*”.

Compositionality: One of the most complex issues faced in the encoding of FDs into sMWN is given by complex linguistic descriptions like <*seagull*> has an orange beak. Complex target concepts such as orange beak cannot be represented as WN synsets, in that in this model synsets are bound to be lexical units.

The solution has been to exploit the notion of *phraset* introduced in MWN for coping with cross-language lexical gaps and with complex ways to express a concept for which a synset already exists (cfr. Bentivogli and Pianta, 2004). In this way, a free combination of words like coltello da pane (the Italian translation for breadknife) is encoded as a phraset {GAP}{coltello_da_pane} linked by the lexical relation *composed-of* to the synsets {coltello} ('knife') and {pane} ('bread'), and by the semantic relation *hypernym* to the synset {coltello} ('knife').

In Lebani and Pianta (2010b) we proposed to exploit the same structure for representing complex descriptions, with the important difference, shown in figure 1, that we represent also the semantic of the modifier (in our example orange),

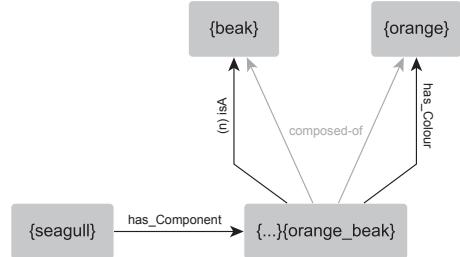


Figure 1: Representation of the FD <*seagull*> has an orange beak

by linking the phraset to the “modifying” synset also with a semantic relation. This allows us to keep track of properties of the described concept that would be otherwise lost.

The set of normalized features: The outcome of the encoding phase has been the insertion into sMWN of 871 normalized descriptions for 5 concepts. On average, every concept received 174.6 descriptions (s.d. 33.44). The results of this encoding confirm that the WN model is apt to represent the kind of commonsense knowledge carried by featural descriptions.

The simplest normalizing procedure has been, as a matter of fact, powerful enough for encoding the vast majority of the collected descriptions. The semantics of 795 normalized FDs (91.3% of the total) could indeed be fully encoded as a semantic relation between two simple synsets. In 137 cases (15.7%) a synset for the focal concept of the description was missing. By exploiting iMWN, 59 equivalent descriptions have been merged together into 29 relations.

The encoding of 71 normalized features required the creation of one or more phrasets, leading to the creation of 76 new phrasets.

In the disambiguation of words we faced an average ambiguity of 3.2 synsets per lemma (s.d. 2.87), and 64 descriptions (7.3% of the sample) have been encoded with more than one relation. In 32 cases a part of the information expressed by the FD, has been discarded. Only 5 raw descriptions were discarded because an efficient way to encode them was not found.

5.2 Modifying the WN model

Even if the bulk of the design of sMWN is the WN model implemented in iMWN, some minor modifications have been necessary to cope with some recurrent problematic kinds of descriptions.

Apart from the exploitation of the phraset structure, we used *relation features*, that is features (labels) associated to relation instances, in

order to refine the semantics of a specific relation-concept pair, along the lines of the proposal advanced by Alonge et al (1998) in the context of the EuroWordNet project.

Negation: In some norms collections, e.g. the McRae database, negative statements are treated as a class on their own, so that FDs like `<bike> doesn't have an engine` and `<chicken> cannot fly` are treated as conveying the same type of information. However, for our purposes, it is important to encode not only that a concept does not possess some property, but also the property it does not possess.

Our solution is the exploitation, in sMWN, of a *negative* operator analogue to that implemented in the EWN database. In this way, a FD like `<chicken> cannot fly` is encoded as a relation of type *is Involved in* between `{chicken}` and `{fly}` and the relation is marked with the *negation* relation feature.

In accordance with the rationale behind the implementation of the *negation* operator in EWN, we noticed that the properties negated by our speakers can be seen as blocking “expected” undesired implications. In our example, indeed, the negated property *fly* is a distinctive property possessed by *birds*, the general category to which the described concept belongs.

Cardinality: This issue affects virtually every work belonging to the feature generation paradigm. Many different solutions have been proposed, but none of them is useful for our purposes. As an example, in Vinson and Vigliocco (2008), descriptions such as *has 4 wheels* are split into the two concepts *4* and *wheels*. However, what is predicated in the pair `<bus> has 4 wheels` cannot be equivalent to what is encoded by associating the concepts *4* and *wheel* to the concept *bus*. McRae and colleagues, on the other side, treated these cases by splitting them in two features (*has wheels* and *has four wheels*), thus introducing some redundancy in their data.

Our proposal is to encode cardinality by means of a *has cardinality* relation feature that specifies the number, numbers or range of numbers of the elements of the set referred to in the description. Accordingly, pairs like `<bus> has 4 wheels` have been encoded as a *has Component* relation, marked with a “*has cardinality:4*” label, holding between the synsets `{bus}` and `{wheel}`. When encoding FDs involving the same synsets with different cardinalities (e.g. `<truck> has wheels`, *may have 4 wheels*, *may have 6 wheels*), we clustered them by

marking the range or set of different cardinalities (in our example, “*has cardinality:4,6*”).

Certainty features: Another common problem for the building of norms collections is the treatment of modifiers like “generally”, “most of the times” and “sometimes”. Standard approaches to feature norms collection remove such expressions in the normalization phase. Also standard WN encoding of semantic relations ignores any kind of qualification of the probability or strength of semantic relations between concepts.

However we think that by ignoring this kind of information an important aspect of lexical meaning gets lost. In the same vein, Boyd-Graber et al (2006) argue for the usefulness of adding to the WN model a characterization of the strength of the relation holding between synsets.

We propose to add a relation feature, called Certainty, representing the intuition of the language speaker about how strong is his/her expectation that a certain relation holds between the instances of two concepts. We distinguish four levels of expectation:

- *True by definition*: the speaker thinks that the relation between two concept instances holds because of how the concepts are conventionally defined; no exceptions are admitted: `<cat> is a feline`.
- *Certain*: the speaker expects the relation to hold unless an anomaly occurs, which needs a causal explanation: `<man> has arms`, `<socks> always come in a couple`.
- *Probable*: the speaker expects the relation to hold most of the times; however if this does not occur it is not perceived as an anomaly. This feature is associated to pairs like `<wardrobe> is typically made of wood`.
- *Possible*: the speaker expects the relation to occur sometimes, but not most of the times. This feature is associated to FDs like: `<wardrobe> can be made of plastic`.

It should be stressed that in the above definitions we are interested in representing a subjective, speaker-oriented, notion of possibility/probability instead of the corresponding formally oriented notions defined in modal logic (Hughes and Cresswell, 1996). Note also that when a FD does not include any type of modifier, it is impossible to decide which of the four classes above it belongs to. Because of this, we represent the Certainty feature only when an explicit linguistic clue allows us to infer a value for it. In all other cases the value of the feature is undefined. We

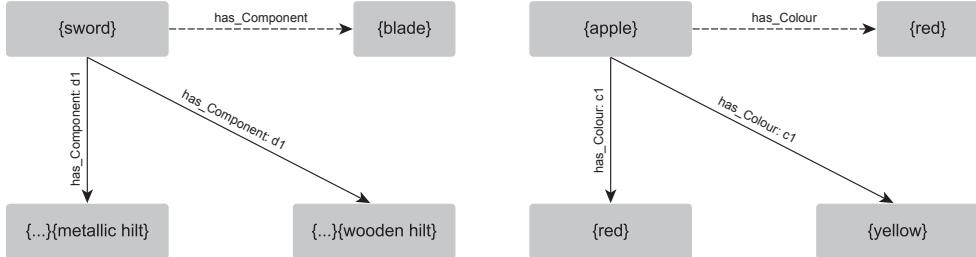


Figure 2: Representation of the FDs `<sword>` has a metallic or wooden hilt (left) and `<apple>` can be red and yellow (right).

reserve for the future the design of further experiments aiming at systematically collecting the value of the certainty feature for all relations, see Nikolova et al (2011).

Conjunction and disjunction: the last set of relation features introduced in sMWN are an implementation of the *conjunction/disjunction* labels introduced in EWN for marking the relation holding between features of the same type that have been predicated of a certain concept.

In sMWN, we set a default value for every semantic relation. As an example, by default the *has Component* descriptions stand in a conjunctive relation, while the *has Colour* ones are disjunctive. As in EWN, moreover, special cases are marked by adding labels to the semantic relations. In this way, the two descriptions `<sword>` can have a wooden hilt and `<sword>` can have a metallic hilt have been encoded in sMWN as shown in figure 2 (left), while figure 2 (right) shows how we encoded conjunctive FDs in a disjunctive environment such as `<apple>` can be red and yellow. In this figure, “ d_i ”/“ c_i ” stands for “disjunction”/“conjunction” and the index points to the other feature(s) standing in a disjunctive/conjunctive relation.

5.3 Comparison with the Kremer Sample

We can get some indications of the goodness of our methodology also from a quick comparison with a parallel sample from the Kremer dataset. For these concepts Kremer and colleagues collected 832 raw descriptions. We annotated their dataset with our feature types, obtaining 231 distinct properties, that is, a mean of 46.2 properties per concept (s.d. 7.95). A chi-square analysis failed to highlight a significant difference in the distribution of raw descriptions across concepts in the two samples ($p > .5$). However, the difference in the average number of features per concept is significant ($W = 25$, $p < .01$).

Moreover, there is a significant difference in the distribution of descriptions in the different feature type classes ($\chi^2 = 75.42$, $df = 9$, $p < .001$). While in our sample there are on average 30.1 description for the 29 represented feature types (s.d. 22.24), in the re-tagged Kremer sample the 23 represented feature types received, on average, 10.04 descriptions (s.d. 9.88).

Our sample, finally, seems to suffer a little less from the problem of disproportionate representation of certain types over others reported by Kremer and Baroni (2011). In the sample from their dataset, indeed, the 6 most frequent relations account for the 62.8% of the whole set of descriptions, while in our sample this measure reduces itself to the 45.1%.

6 Conclusions and future work

In this paper we presented our reflections and preliminary work for the creation of a WordNet that can be exploited for therapeutic purposes. Even if created with a specific applicative use in mind, we conceived this resource as to be able to represent every kind of knowledge that can be associated with a concrete concept.

By modifying the WN model, we've been able to represent a subset of the descriptions we collected from 60 Italian speakers. Even if we concentrated only on a subset of our collection, we feel safe to claim that we demonstrated that it is possible to represent in a WN-like resource all the semantic information that can be collected through a description elicitation experiment.

There are, still, many steps left to go. We are currently mapping all the remaining features of our collection and we are testing the reliability and the intuitiveness of our feature type classification. Given that building a norms collection is a time consuming task (McRae and colleagues begun working on their collection in the 90s), an issue that we will face in the immediate future is

how to automatically mine and annotate the commonsense knowledge to encode into WN.

Furthermore, being our resource based on a multilingual version of WN, i.e. MWN, another issue we're going to pursue is the evaluation of the portability of the information we elicited from our participants to languages other than Italian.

Acknowledgments

This work has been partly supported by the *Provincia autonoma di Trento* and the *Fondazione Cassa di Risparmio di Trento e Rovereto*.

References

- Antonietta Alonge, Nicoletta Calzolari, Piek Vossen, Laura Bloksma, Irene Castellon, Maria A. Martí and Wim Peters. 1998. The linguistic design of the EuroWordNet database. *Computer and the Humanities*, 32 (2,3): 91-115.
- Eduard Barbu and Massimo Poesio. 2008. A comparison of Wordnet and Feature Norms. *Proceedings of GWC 2008*: 56-73.
- Luisa Bentivogli and Emanuele Pianta. 2004. Extending WordNet with Syntagmatic Information. *Proceedings of GWC 2004*: 47-53.
- Luisa Bentivogli, Pamela Forner, Bernardo Magnini and Emanuele Pianta. 2004. Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. *Proceedings of COLING 2004*: 101-108.
- Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson and Robert Schapire. 2006. Adding Dense, Weighted, Connections to WordNet. *Proceedings of GWC 2006*: 29-36.
- Simon De Deyne, Steven Verheyen, Eef Amel, Wolf Vanpaemel, Matthew J. Dry, Wouter Voorspoels and Gert Storm. 2008. Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, 40 (4): 1030-1048.
- Christiane Fellbaum. 1998a. *WordNet. An electronic lexical database*. The MIT Press. Cambridge, MA.
- Christiane Fellbaum. 1998b. A Semantic Network of English Verbs. In C. Fellbaum (ed.) *WordNet. An electronic lexical database*. The MIT Press. Cambridge, MA: 69-104.
- Peter Garrard, Matthew A. Lambon Ralph, John R. Hodges and Karalyn Patterson. 2001. Prototypicality, distinctiveness and intercorrelation: analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology* 18: 125-174.
- George E. Hughes and Max J. Cresswell. 1996. A New Introduction to Modal Logic. Routledge. London and New York.
- Rubén Izquierdo, Armando Suárez and German Rígau. 2007. Exploring the automatic selection of basic level concepts. *Proceedings of RANL 2007*: 298-302.
- Gerhard Kremer and Marco Baroni. 2011. A set of semantic norms for German and Italian. *Behavior Research Methods*, 43 (1): 97-109.
- Gianluca E. Lebani and Emanuele Pianta. 2010a. A Feature Type Classification for Therapeutic Purposes: a preliminary evaluation with non expert speakers. *Proceedings of ACL-LAW IV*: 157-161.
- Gianluca E. Lebani and Emanuele Pianta. 2010b. Exploiting Lexical Resources for Therapeutic Purposes: the Case of WordNet and STaRS.sys. *Proceedings of COLING-CogALex-II*: 94-101.
- Ken McRae, George S. Cree, Mark S. Seidenberg and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments & Computers*, 37 (4): 547-559.
- George A. Miller. 1998. Nouns in WordNet. In C. Fellbaum (ed.) *WordNet. An electronic lexical database*. The MIT Press. Cambridge, MA: 23-46.
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller. 1990. Introduction to WordNet. An On-line Lexical Database. *International Journal of Lexicography*, 3 (4): 235-244.
- Gregory L. Murphy. 2002. *The big book of concepts*. The MIT Press, Cambridge, MA.
- Sonya Nikolova, Jordan Boyd-Graber and Perry R. Cook. 2009. The design of ViVA: a mixed-initiative visual vocabulary for aphasia. *Proceedings of CHI'09*: 415-420.
- Sonya S. Nikolova, Jordan Boyd-Graber and Christiane Fellbaum. 2011. Collecting semantic similarity ratings to connect concepts in assistive communication tools. In A. Mehler et al (eds.) *Modeling, Learning and Processing of Text Technological Data Structures*. Springer. Berlin and Heidelberg: 81-93.
- Brian Petheram. 2004. Special Issue on Computers and Aphasia. *Aphasiology*, 18 (3): 187-282.
- Emanuele Pianta, Luisa Bentivogli and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. *Proceedings of GWC 2002*: 292-302.
- Dan Tufiş. 2004 ed. Special Issue on BalkaNet. *Romanian Journal on Information Science and Technology*, 7 (1,2): 1-248.
- David P. Vinson and Gabriella Vigliocco. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40 (1): 183-190.

Visual Study of Estonian Wordnet using Bipartite Graphs

and Minimal Crossing algorithm

Ahti Lohk

Tallinn University of Technology
Tallinn, Estonia
Ahti.Lohk@ttu.ee

Kadri Vare

University of Tartu
Tartu, Estonia
Kadri.Vare@ut.ee

Leo Vohandu

Tallinn University of Technology
Tallinn, Estonia
Leo.Vohandu@ttu.ee

Abstract

Over the last years the Estonian Wordnet (EstWN) has grown quite large – currently there are around 45 000 concepts present. The work is done by different people over many years and that is the reason why the EstWN needs constantly to be checked for errors and other non suitable concepts, semantic relations etc. In this paper we describe a visualization algorithm (diagnostics tool), which indicates to possible problems in EstWN¹.

1 Introduction

At present the main goal is to increase EstWN with new concepts and enrich EstWN with different kinds of semantic relations. But at the same time it is necessary to check and to correct the concepts already present (Kerner, 2010; Orav et al 2008).

One of the issues is that originally in EstWN the aim was to avoid (if possible) multiple hyperonyms (Vider, 2001). In EstWN there are currently 1 117 concepts (synsets) with two hyperonyms, 134 concepts with three or more hyperonyms and the concept which has the most hyperonyms – 9 – is 'alkydecolour'. Of course, some synsets might have multiple hyperonyms, also the EuroWordNet² format allows this. Often however the situation where a synset has multiple hyperonyms happens if the lexicographer adds a new and more precise hyperonym in editing process but forgets to delete the old one; or the lexicographer can't decide which hyperonym fits better. Usually the multiple hyperonyms are

all somehow related to its hyponym but only one of them is the most suitable one and others can be corrected into some other semantic relations. Similarly it is important to detect and to remove false hyperonyms even if the synset has only one hyperonym.

For a lexicographer it is much easier to have data visualized in some way, so the errors can be more easily detected (similar work has been done for Chinese WordNet, see Xu et al 2008).

In what follows we take a more general viewpoint. We allow all relations between synsets which are standard in wordnet.

Wordnet can be formally described as a collection of synsets (concepts). On that set there are defined different relations (hyperonymy, hyponomy, near synonymy etc) between synsets or even words (depends on how we look at this system). As the number of words in wordnet is big enough and any word can have many different relations to other words it is easy to understand that we have to do with a very complex system of connections which are hard to handle and to check.

For a deeper interactive study of any wordnet system we did build a special workbench enabling to present the finite separable subsets of a relation in a formal way (bipartite graphs) and then in a visualization mode which makes it easy for a lexicographer to estimate the quality of separated natural subsets (closed sets).

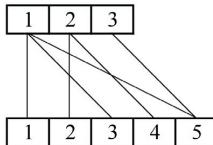
2 Method

We will explain our method's main idea with a small artificial example. Let us have a small separated subset presented as a matrix:

¹ <http://www.cl.ut.ee/ressursid/teksaurus/>

² <http://www illc.uva.nl/EuroWordNet/>

1	1	0	0
2	0	1	0
3	1	0	0
4	0	1	0
5	1	0	1
	1	2	3

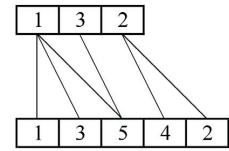


In the rows of that table we have synsets and in columns hyperonyms. On the right side of that figure we have presented the same data as a bipartite graph where all column numbers are positioned on the upper line and all rows on the lower line. Every connecting line on the right side has been drawn between every “1”-s column and row number.

As we see there exist a lot of line crossings even in our very small example. It is possible to reorder the rows and columns of that table into optimal positions so that the number of line crossings would be minimal possible. If there is

full order then there will be no crossings of lines! In general case this crossing number minimization is a NP-complete task. We are using the idea of S. Niermann's (2005) evolutionary algorithm to minimize the number of line crossings. In our example the optimal result will be:

1	1	0	0
3	1	0	0
5	1	1	0
4	0	0	1
2	0	0	1
	1	3	2



As we can see there are no crossings and all connections are separated into two classes. We have got a nice and natural ordering for rows and columns. With that kind of picture the relations between words (synsets) are easier to see and understand. We will present real cases from EstWN later.

hyperonyms	1	kosjaminek_1_n		
	2	kosjarcis_1_n		
	3	kosjaskäik_1_n		
	4	kosjasõit_1_n		
	5	kosjatee_1_n		
	6	rats_1_n		
	7	ratsionaliseerimisetepanek_1_n		
	8	armulaud_1_n		
	9	paaripanek_2_n		
	10	religioosne rituaal_1_n		
	11	riitus_1_n		
	12	viljakusrituaal_1_n		
	13	haigete salvimine_1_n		
	14	haigete võidmine_1_n		
	15	konfirmatsioon_1_n		
	16	leer_1_n		
	17	leerikool_1_n		
	18	ordinatsioon_1_n		
	19	peakool_1_n		
	20	piht_3_n		
	21	pihtimine_1_n		
	22	ristimine_1_n		
	23	viimne salvimine_1_n		
	24	viimne võidmine_1_n		
hyperonyms	1	ettepanek_2_n		
	2	rituaal_1_n		
	3	sakrament_1_n		
	4	võidmine_1_n		

Table 1: Closed set

We have to add that for the whole wordnet to use crossing minimization directly would be practically impossible. To make it easier for the computer and also for the lexicographer we first separate from the wordnet all connected components (closed sets) for the chosen relation (Table 1). We use a standard algorithm from D. Knuth's encyclopedia (1968) which creates all connected components very quickly. For example using *hyperonymy* relation on all nouns for EstWN it takes only 2.5 minutes on a standard laptop to get all 4 854 closed sets.

3 Examples of closed sets in EstWN

It is possible to determine closed sets in EstWN considering *hyperonymy-hyponymy* relations of verbs and nouns. For adjectives and adverbs we decided that the best relation to create closed sets was *near synonymy*. Altogether it was possible to determine 4 854 closed sets for nouns, 1 178 sets for verbs, 478 for adjectives and 331 for adverbs.

Each closed set should consist of word senses which are semantically related. It is also possible to create closed sets by using other semantic relations present in EstWN. The more members the closed set has the higher is the chance that there are also non suitable words. For example the largest closed set of verbs contains 481 words and *hyperonyms* in this set are 'to inform', 'to sound', 'to move', 'to hit'. Many *hyponyms* in this set are onomatopoetic words – they can denote movement and sound at the same time and this explains why *hyperonyms* such as 'sound', 'move' and 'hit' and their *hyponyms* might belong to the same set. But the verb 'to inform' is connected to this closed set only because one of the *hyponym* 'to knock' has also a *hyperonym* 'to hit' ('knock' in Estonian means also to 'to hint', 'to narc').

In Table 1 there is an example of a closed set for nouns. It can be seen that the word *rationaliseerimisettepanek* ('labour-saving proposal') does not belong to this semantic field (this semantic field can be named 'different kinds of rituals' for example). This closed set can be seen as a bipartite graph where words with gray background are *hyperonyms* to words with white background.

In Table 1, only the first synset literal is presented as hyperonym. That kind of approach has been chosen because of two reasons: on the one hand EstWN web application uses this type of format and on the other hand before applying the

crossing number minimization we separate all connected components (closed sets) as it was mentioned before. If the components are separated then a lexicographer can have an overview of how many hyperonym-synsets are connected with this closed set and that hopefully leads his/her attention to the potential problems.

After the closed sets are formed it is possible to apply the minimal crossing algorithm to a chosen set.

Minimal crossing in Figure 1 is created based on the data in Table 1. While applying Minimal Crossing it is not important if the connected objects in the matrix are word-word (as in Figure 1) or synset-synset (as in Figure 2).

4 Possible ways of using the Minimal Crossing method

There are different possibilities where the method of Minimal Crossing can be helpful for detecting to errors and problematic cases.

As next we propose four points which illustrate the possible benefits of the Minimal Crossing method.

- It is possible to different kinds of semantic relations to create closed sets. For the closed set of the nouns in Figure 1 we used *hyperonymy* as the connecting relation and in Figure 2 the closed sets are done by connecting adjective synsets which are *near synonyms*.
- It is possible to detect subgroups. For example in figure 1 one set of words is subgroup of another group. A lexicographer can correct relations between these synsets. In this example the synset *võidmine* ('unction') should be probably hyponym to *sakrament* ('sacrament').
- It is possible to detect wrong and missing semantic relations. For example it is strange that words *rationaliseerimisettepanek* ('labour-saving proposal') and *kosjakäik* ('a visit to bride's house to make a marriage proposal') belong to same closed set. Both these synsets share a hyperonym *ettepanek* ('proposal'), but *kosjakäik* should be connected to *ettepanek* ('proposal') by *is_involved* relation and the hyperonym to *kosjakäik* should be 'ritual' instead.
- It is possible to detect missing senses. In the adjective closed set example in figure 2 words '*taevalik*' ('heavenly') and *perversne* ('perverse') belong to the same closed set (are near synonyms of near synonyms). By

examining the senses and relations of these words and synsets it appears that one word – *ebaloomulik* ('abnormal') – has been marked as a near synonym but the sense is an inaccurate one. The correct sense of *ebaloomulik*

('abnormal') is missing altogether from EstWN.

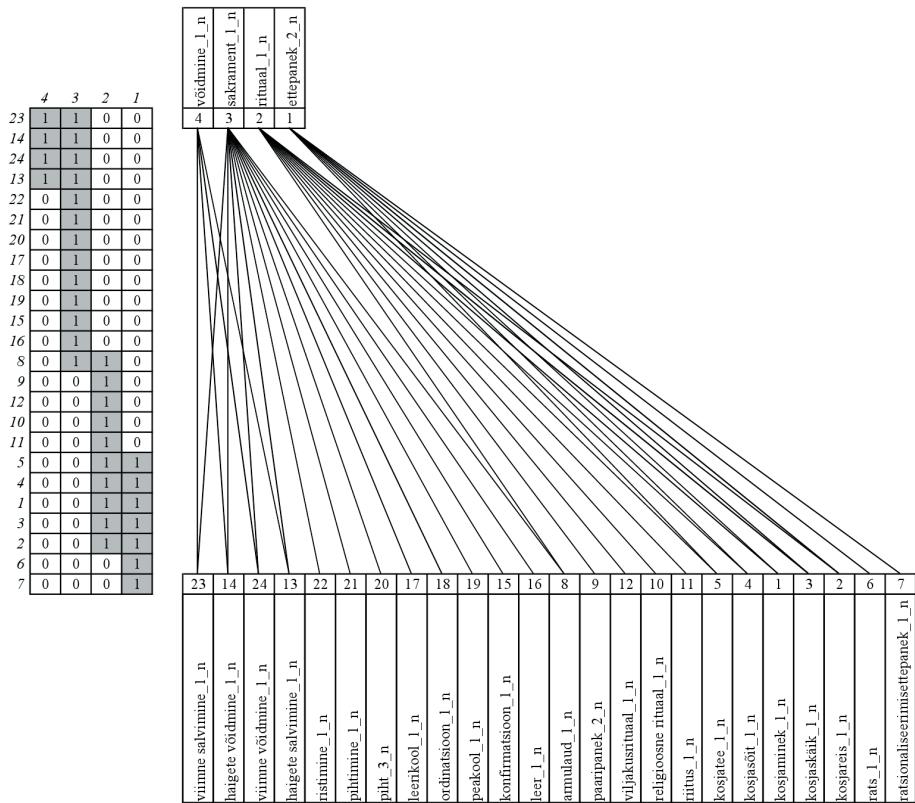


Figure 1. Bipartite graph and minimal crossing with hyperonyms.

6	1	1	0	0	0	0
2	0	1	1	0	0	0
1	0	1	1	1	0	0
5	0	0	0	1	0	0
4	0	0	0	1	1	1
3	0	0	0	0	0	1
	4	5	1	3	6	2

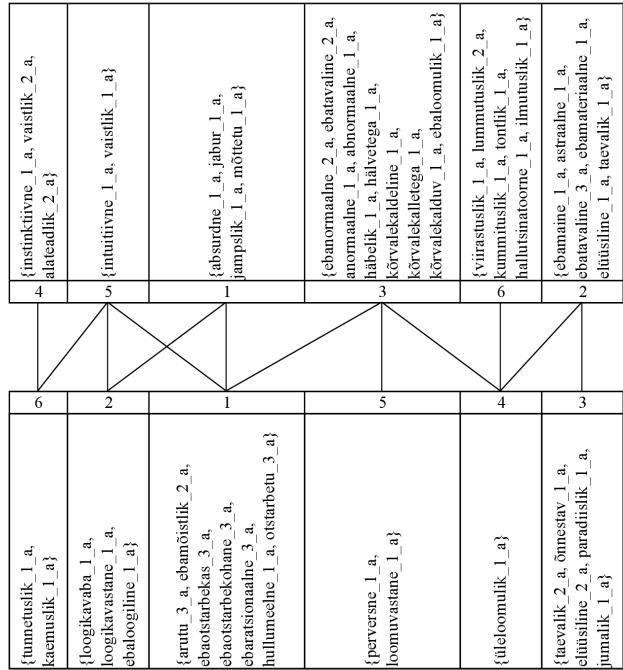


Figure 2. Bipartite graph and minimal crossing with near synonyms.

Synsets (upper part): 4-instinctive, instinctual, 5-intuitive, non-rational, 1-absurd, non-reasonable, 3-abnormal, 6-ghostlike, apparitional, 2-nonmaterial, spiritual;

Synsets (lower part): 6-glandular, intuitive, 2- un-logical, illogical, 1-unwise, 5-perverse, deviant, 4-supernatural, 3-divine, dysian, inspired, sublime.

5 Conclusion and future work

The created workbench has already proven its usefulness in lexicographer's practical work.

Lexicographer doesn't need programmer's help to work with the system.

Also, this algorithm can be used to cluster concepts in EstWN.

As next we will use the tool to study the inners of Princeton Wordnet (Miller, 1990).

References

- Kerner K. Orav H. Parm S. Growth and Revision of Estonian WordNet. In: *Principles, Construction and Application of Multilingual Wordnets. Proceeding of the 5th Global Wordnet Conference: 5th Global Wordnet Conference; Mumbai, India; 31.01-4.02.2010. (Edit.) Bhattacharyya, P.; Fellbaum, Ch.; Vossen, P.* Mumbai, India: Narosa Publishing House, 2010, pp 198-202.
- Knuth D.E. 1968, *Fundamental Algorithms, vol. 1 of Art of Computer Programming* (Reading, MA, Addison-Wesley), §2.3.3.
- Vider K. Eesti keele tesaurus - teooria ja tegelikkus Leksikograafiaseminar "Sõna tänapäeva maailmas" Leksikograafinen seminaari "Sanat nykymaailmassa". Ettekannete kogunik. Toim. M. Langemets. Eesti Keele Instituudi toimetised 9. Tallinn 2001 lk 134-156.
- Niermann S. Optimizing the Ordering of Tables With Evolutionary Computation. *The American Statistician* 2005, 59(1):41-46.
- Miller, Georg; Beckwith, R; Fellbaum, Ch; Gross, D; Miller, K.J 1990. Introduction to WordNet: An Online Lexical database. – *International Journal of Lexicography* 3.
- Orav, H.; Vider, K.; Kahusk, N.; Parm, S. (2008). Estonian WordNet: Nowadays. In: *Proceedings of the Fourth Global WordNet Conference: GWC 2008; Szeged, Hungary; January 22-25, 2008.*

(Edit.) Tanacs, A.; Csendes, D.; Vincze, V.; Fellbaum, Ch.; Vossen, P., 2007, 334 - 338.

Xu, Ming-Wei, Jia-Fei Hong, Shu-Kai Hsieh, and Chu-Ren Huang. 2008. CWN-Viz : Semantic Relation Visualization in Chinese WordNet. *In: Proceedings of the Fourth Global WordNet Conference: GWC 2008; Szeged, Hungary; January 22-25, 2008.* (Edit.) Tanacs, A.; Csendes, D.; Vincze, V.; Fellbaum, Ch.; Vossen, P., 2007, 506-519.

Rethinking WordNet's Domains

Xiaojuan Ma

Carnegie Mellon University
Pittsburgh, PA, United States.

xm@cs.cmu.edu

Christiane Fellbaum

Princeton University
Princeton, NJ, United States
fellbaum@princeton.edu

Abstract

WordNet's current domain structure is incomplete and semantically heterogeneous. Based on our examination of the domain taxonomies of a number of resources as well as a critical evaluation of the WordNet Domains Project, we suggest principles to semi-automatically construct a comprehensive domain ontology for WordNet and propagate domain labels throughout the network in ways that would benefit automatic word sense disambiguation.

1 Introduction

WordNet (Fellbaum, 1998) is a much-used resource in NLP applications, where it supports word sense disambiguation with rich lexical-semantic information, including synonyms, hyper- and hyponyms, meronyms, antonyms, and definitions (“glosses”). Because WordNet aims to represent human knowledge about concepts and the words expressing them in a structured fashion, it is often called an “ontology.”

WordNet 3.1 includes more than 117,000 synsets, ranging from very general concepts like “food” to highly specific ones like “tipsy_cake” and “baked_Alaska.” All synsets are ultimately linked to a single top node “entity” and thus a path exists between any pair of synsets.

WordNet's rich coverage can obscure a coarser view of the lexical and conceptual inventory of English speakers. But for a variety of reasons, users may want to view the lexicon at neither a too abstract nor a too fine-grained level. One practical incentive is that a broader categorization allows one to divide WordNet into topical subnets, thus reducing the search space among related words and concepts. Such subnets would moreover interlink synsets across parts of speech (POS), in addition to WordNet's largely within-POS connections. Construction of a domain organization is further motivated by the intrinsic interest in a fundamental question going back to Aristotle's *Metaphysics*: (how) can human knowledge be categorized in an exhaustive and non-overlapping fashion? This question faced the

creators of WordNet as much as all of its related predecessors such as Roget's thesaurus and numerous encyclopedic and library classification efforts.

The remainder of this paper provides a critical discussion of the current “domain” structure of WordNet as well as related alternative proposals. Section 3 provides a summary of a wide range of domain-related classification systems and their characteristics. In Section 4 we propose some general principles and guidelines for constructing a domain ontology for WordNet.

2 WordNet as an Ontology

Following WordNet's own definition, we take an ontology to be a “rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations.” WordNet's ambition to cover most of the English lexicon (and thus, by extension, most of the concepts denoted by the lexical entries) required its lexicographers to classify all domains of human knowledge. This momentous task was approached with the “divide and conquer” strategy, whereby words from different lexical categories were treated separately as sublexicons and a broad classification was undertaken for each of them (Fellbaum, 1998; Miller, 1998). This classification resulted in 45 “lexicographer files¹”. The lexicographer files provide a rough semantic classification, and the assignment of different senses of a polysemous word to different lexicographer files has supported word sense disambiguation, e.g. (Siegel, 1998).

However, the lexicographer files do not carve up the lexicon in a consistent and non-overlapping way. While categories like “animal,” “plant,” “food” and “artifact” seem fairly straightforward, “possession” and “object” are more difficult. One reason is the relation between the category and its members is not the same here. Moreover, all artifacts *can* be possessions, but since this is not *necessarily* the case, cross-

¹ <http://wordnet.princeton.edu/man/lexnames.5WN.html>.

classification seems wrong. “Object” is defined as “a tangible and visible entity; an entity that can cast a shadow” (as opposed to an abstract entity); clearly, this category overlaps largely with artifact. But only in very few cases are noun synsets assigned to more than one superordinate. Note further that “state” and “process” pick out entities with a temporal meaning component but no other semantic commonalities; similar, the verb file “stative” includes verbs with shared aspectual properties that are otherwise semantically heterogeneous (Fellbaum, 1990). Finally, adjectives are not classified into broader semantic categories but are grouped into thousands of unconnected “clusters”.

2.1 Domain Categories in WordNet 3.0

The lexicographer files served primarily to facilitate the huge task facing WordNet’s lexicographers. As the examples above indicate, they are not well suited to classify the lexicon into categories that are neither too abstract nor too specific and that may significantly improve word sense discrimination. Beginning with version 2.0, bidirectional, cross-POS pointers were added to WordNet that connected “domain” synsets like {baseball, baseball_game} to noun synsets like {inning, frame}, verb synsets like {retire, strike_out} and adjective synsets such as {outside, away}. While these links provided additional semantic information, the “domain” pointer is fraught with problems. To begin with, it includes three very different kinds of relations: domain *categories* (e.g., {Chinese brown sauce} ↔ {cooking}), domain *regions* (e.g., {Chinese brown sauce} ↔ {China}), and domain *usage* (e.g., {in_the_lurch} ↔ {idiom}). Here we consider only domain categories. We distinguish *domain categories* (synsets like {cooking, cookery}) and *domain members*, synsets like {Chinese brown sauce} and {blanch, parboil} that are linked to the categories.

One problem with the present WordNet domain system is sparsity. A total of 6428 synsets (4096 nouns, 1224 verbs, 1073 adjectives, and 35 adverbs) are linked to 438 domain categories in WordNet 3.0, covering only 5% of all the synsets in the database. Among all domain categories, 356 include noun synsets, 165 have verb synset members, 131 include adjective synsets, and 17 have adverb members. In a word, the current domain categories are still too few to be truly useful for word sense discrimination.

A second problem with the current domain organization is their shallow hierarchical structure,

which never exceeds three levels (e.g., {law} → {criminal law} → {crime}). Only 51 out of 438 domain members are themselves members of domain categories.

Third, the semantic distance of the domain categories at the same level to the very top of WordNet is very different. In other words, synsets that were chosen as domain categories are not uniform in terms of their degree of generality (or specificity) (e.g. {arabian nights’ entertainment} is a very specific domain category, while {art} is rather abstract).

Fourth, a large number of domain categories are not assigned with any domain information. For instance, the domain {mathematics} is a member of {science}, but {basketball} is not associated with {sport} in the category hierarchy.

Finally, there are loops within the domain hierarchy: {computer} and {computer science} are the domain category of each other.

In sum, the domain category system in WordNet 3.0 needs significant improvement before it can reliably support NLP applications.

2.2 The WordNet Domains Project

The WordNet Domains Project (WNDomains)² constitutes an independent attempt to incorporate domain information into WordNet. Bentivogli et al. (2004) and Magnini and Cavaglia (2000) augmented WordNet version 1.6 with domain tags derived from subject field codes from the Dewey Decimal Classification (cf. Section 3). The latest published version of WNDomains includes 168 domain labels in a hierarchical structure; 45 terms are considered the basic domains. Six top-level domains are “doctrines,” “free_time,” “applied_science,” “pure_science,” “social_science,” and “factotum” for generic and unclassified synsets.

Over 115,000 synsets were annotated with WNDomains tags through an iterative two-step approach: (1) seeding the tags manually in a small number of high-level synsets; (2) automatic propagation of the tags to related synsets (hyponyms, troponyms, meronyms, antonyms and pertainyms). The results of each iteration were evaluated using a text classification task.

2.3 Comparison between WordNet 3.0 Domain Categories and WNDomains

WNDomains has not been mapped to more recent versions of WordNet. We performed a sense mapping between WordNet 1.6 and 3.0 to extract

² <http://wndomains.fbk.eu/>.

the overlapping synsets that are labeled by both domain categories and WNDomains (Table 1). Out of the 6,428 synsets annotated in WordNet 3.0, only 4,884 were tagged in WNDomains, and only 1,927 synsets have the same category labels in both resources (e.g. accuracy%1:07:03:: in Table 1). For the remaining synsets, the domain category was either more specific (e.g. xy%1:08:00::) or more generic (e.g. water_ski%2:38:00::) than the corresponding WNDomains tag, or the two domain classifications disagree (e.g. nucleosynthesis%1:22:00::). Note that a synset may be a member of several WordNet categories (e.g. fault %1:04:01::) -- an indicator that a common supercategory would be appropriate.

WordNet 3.0 Synset	WordNet 3.0 Domain Category	WNDomains
accuracy%1:07:03::	mathematics%1:09:00::	mathematics
copyrighted%5:00:00:proprietary:00	music%1:10:00::, literature%1:10:00::	factotum
fault%1:04:01::	squash%1:04:00::, tennis%1:04:00::, badminton%1:04:00::	table_tennis, tennis
nucleosynthesis%1:22:00::	astronomy%1:09:00::	chemistry
water_ski%2:38:00::	sport%1:04:00::	skiing
world_premiere%1:10:00::	music%1:10:00::	racing, sociology, telecommunication, theatre
xy%1:08:00::	genetics%1:09:00::	biology

Table 1: Examples of domain categories and WNDomains tags for synsets in WordNet 3.0.

The comparison also reveals several important differences between WordNet 3.0 domain categories and WNDomains. First, the Princeton WordNet domain categories are synsets while WNDomains tags are words that are not disambiguated. For instance, one of the WNDomains tag is “play,” and there is no clue on its part of speech or actual meaning when used in annotation. Out of the 168 WNDomains tags, only 83 of them share the same word form as one of the WordNet 3.0 domain categories.

Second, WNDomains created a special domain called “factotum” to classify generic synsets like {man} that are highly polysemous, as well as synsets that appear frequently in various contexts and are considered “stop” senses, such as numbers and colors. Out of 115,424 synsets labeled in WordNet 1.6, about 41,000 (35.52%)

were assigned to the “factotum” tag. This category seems far too broad and too heterogeneous. Moreover, the label “factotum” is infelicitous, as this word denotes a servant or assistant, clearly not a concept that applies to the members of this domain.

Third, synsets from different parts of speech were necessarily annotated separately in WNDomains since the labels were propagated using WordNet relations, which link synsets within the same lexical category only. In other words, even though the same set of labels were used to annotate nouns, verbs and adjectives, the labels had to be seeded independently and the automatic inheritance procedure could only be carried out within the hierarchies for each part of speech. By contrast, WordNet domain categories spread the domain information across different word classes.

In sum, while WNDomains provides some useful domain information for WordNet, it fails to fully utilize the WordNet structure and update its coverage.

3 Other Domain Taxonomy Resources

WNDomains was one of the first attempts to augment WordNet with a domain hierarchy inferred from other resources (Magnini and Cavaglia, 2000), which researchers have been using for various tasks in Natural Language Processing. In this section, we present a review of existing domain taxonomies, including their history, characteristic features and examples. For comparison, we list the top levels and the extended lower levels under the domain “social / society / social science” of each taxonomy.

3.1 Library Classification Systems and Subject Headings

Library classification systems are schemes to code and shelve books, videos, electronic resources, etc. to support cataloguing, storage and retrieval in an intuitive yet logical manner. The materials are hierarchically organized based on the subject they cover. A system usually consists of classification (i.e. a system to organize knowledge) and notation (i.e. a symbol system to index the classes identified in classification). The following universal library classification systems aim at comprehensive subject coverage:

- The Dewey Decimal Classification (DDC)³ uses fields of study / disciplines

³ <http://www.oclc.org/dewey/>.

for classification. DDC is divided into 10 main classes (Table 2), each class has 10 divisions, and each division is further broken down to 10 sections.

- The Library of Congress Classification (LCC)⁴ has 21 basic classes, each indexed by a single letter. Most of these are further broken into more specific divisions (subclasses). Each division contains a loose hierarchy of topics pertaining to the subclass.
- Universal Decimal Classification (UDC)⁵ categorizes human knowledge into 10 classes; each class is divided into logical parts (divisions), and each division is further decomposed. The core version of UDC contains more than 68,000 domain labels.

Despite the differences in their notation and classification systems, DDC, LCC and UDC arrange all branches of human knowledge in a hierarchical tree structure based on the subjects of learning (i.e. disciplines), such as “social science,” “natural science,” and “applied science.”

Dewey Decimal Classification
000 Computer science, information & general work
100 Philosophy & Psychology
200 Religion
300 Social science
300 Social science
310 General statistics
320 Political science
330 Economics
340 Law
350 Public administration
360 Social services; association
370 Education
380 Commerce; communications; transport
390 Customs, etiquette, folklore
400 Language
500 Science
600 Technology
700 Art & recreation
800 Literature
900 History & geography

Table 2: Main classes of DDC.

Library catalogs additionally rely on subject headings. A subject heading comprises several aspects, often with topical, geographical, and chronological information included. Topical

headings in particular are words or phrases from a controlled vocabulary to describe the subject matter of specific material. Subject headings support the keyword-based search in library collections.

- The Library of Congress Subject Headings (LCSH) system is a thesaurus developed and maintained by the U. S. Library of Congress to represent bibliographical records.
- The Sears List of Subject Headings (Miller and Sears, 2004), based on the model of LCSH, addresses the needs of small-to-medium-sized libraries. It is continuously updated (the 20th edition was published in 2010) to stay current.

Unlike library classification systems, subject headings were built on topics rather than disciplines and were organized hierarchically via relations such as “Use (USE),” “Used For (UF),” “Broader Topic (BT),” “Related Topic (RT),” “See Also (SA),” and “Narrower Topic (NT).”

3.2 Dictionary Subject Field Codes (SFCs)

In library systems, domain information is attached to physical or digital documents and materials, whereas in dictionaries, semantic codes called Subject Field Codes (SFCs) are assigned to lexical entries. We examined the SFC organization in these dictionaries:

- The Longman Dictionary of Contemporary English (LDOCE) (Longman, 1978) contains 35,899 entries and 53,838 word senses. SFCs (124 major fields/classes and 250 sub-fields/divisions) were manually assigned to 95% of the senses. However, over half of the senses were coded “XX” for the General class, and only 45% of the senses have more domain-informative codes. The SFCs information in LDOCE’s machine-readable version has been used natural language processing tasks (Guthrie et al., 1991; Liddy et al., 1993; Walker and Amsler, 1986; Wilks and Stevenson, 1998). Longman Lexicon of Contemporary English (LLOCCE) is a smaller dictionary extracted from LDOCE with 14 major codes (classes), 127 group codes (divisions), and 2441 set codes (sections).

- Collins English Dictionary (CED) covers 500,000 words. Similar to the subject heading in the library catalog systems,

⁴ <http://www.loc.gov/catdir/cpso/lcc.html>.

⁵ <http://www.udcc.org/>.

CED associates word senses with categories that denote subject-field (SFCs), time, nationality, usage, appropriateness, connotation, or trademarks.

- The Oxford English Dictionary (OED)⁶ reviews the meanings and history of 600,000 English words across all English-speaking countries/regions and is the world's largest dictionary. The OED has 230,000 categories in its classification system, which are hierarchically grouped into three main sections: "The External World," "The Mind," and "Society." OED's categories are more in line with topics than with disciplines (Table 3).

Oxford English Dictionary
• The external world (the world)
• The mental world (the mind)
• The social world (society)
◦ The community
◦ Inhabiting or dwelling
◦ Armed hostility
◦ Authority
◦ Morality
◦ Education
◦ Religion
◦ Communication
◦ Travel
◦ Occupation
◦ Leisure

Table 3: The first two level of subject field codes of OED.

- The UCREL Semantic Analysis System (USAS) contains tags for parts of speech and semantic fields, and aims to facilitate dictionary-based content analysis (Archer et al., 2004). It has 21 major classes and 232 category labels organized in a three-level taxonomy. Senses within the same semantic space are assigned to the same category. The top levels of USAS tags were largely derived from LLOCE's SFCs. Some classes were directly inherited ("general and abstract terms," "entertainment, sports and games"); others are transformed (the "space and time" class in LLOCE is decomposed into two classes in USAS, "time" and "the world and our environment"). Fundamentally a machine readable lexicon with domain tagging, USAS shows a certain similarity to the WNDomains project.

⁶ <http://www.oed.com/view/th/class/1>.

The online version of LDOCE provides topic information⁷ for browsing and searching, and the topics are loosely connected through "related topics." For example, topics related to "society" include "anthropology," "nationality and race," "crime and law" and "youth."

Overall, SFCs in dictionaries are more topic-oriented other than discipline-oriented. Both the traditional semantic relations and topic-relatedness are used to construct the ontology.

3.3 Encyclopedias

Encyclopedias are reference works that provide information for all branches of knowledge.

- The general-coverage Encyclopædia Britannica contains a one-volume Propædia presenting a logic outline of human knowledge. Propædia's hierarchical (Table 4) framework consists of 10 parts (classes), 41 divisions, and 167 sections.
- Wikipedia is a web-based multilingual encyclopedia openly edited by contributors from all over the world. Created in 2001, the English Wikipedia alone currently includes over 3,735,000 articles and attracts four million visitors monthly. To organize the huge amount of information on all areas of human knowledge, articles on related topics are grouped into content categories⁸.

Propædia
1. Matter and energy
2. The Earth
3. Life on earth
4. Human life
5. Human society
5.1 Social groups: peoples and cultures
5.2 Social organization and social change
5.3 Production, distribution, & utilization of Wealth
5.4 Politics and government
5.5 Law
5.6 Education
6. Art
7. Technology
8. Religion
9. The History of Mankind
10. The Branches of Knowledge

Table 4: Main parts (and society divisions) of Propædia.

In general, Propædia's classification is more topic-based, while Wikipedia's categorization scheme is a mix of topics and disciplines.

⁷ <http://www.ldoceonline.com/browse/topics.html>.

⁸ http://en.wikipedia.org/wiki/Portal:Contents/Categorical_index.

3.4 Corpus-based Taxonomies

A corpus is a collection of written or spoken text. Corpus searches are often facilitated by classification systems.

- The British National Corpus (BNC) is a balanced corpus representing British English usage in the late 20th century. The BNC labels contents with subject field (Table 5) information.

British National Corpus (Written)	
• Imaginative/creative	
• Informative	<ul style="list-style-type: none"> ○ Applied science ○ Arts ○ Belief and thought ○ Commerce and finance ○ Leisure ○ Natural/pure sciences ○ Social sciences <ul style="list-style-type: none"> ▪ Psychology ▪ Sociology ▪ Linguistics ▪ Social work ○ World affairs
• Unclassified	

Table 5: Subject fields for written samples of BNC.

Although not corpora in conventional sense, web directories provide links to a considerable list of web sites and web pages on different topics in interconnected category hierarchies.

- Yahoo! Directory⁹ was the first to provide web directory service. The Open Directory Project (ODP)¹⁰ is the largest directory over the Internet.
- Taxonomy Warehouse¹¹, an online directory, specializes in taxonomies constructed for different fields, such as the ACM Computing Classification System. The website organizes its collection of taxonomies according to their subject domains in a two-level tree structure.

Corpora generally organize their data with a mix of topics and disciplines. For web directories in particular, frequency/popularity of the topics is an important criterion to determine which categories should be placed at a higher level; for example, “shopping” and “kids and teens” are two main categories in ODP. This classification differs from the others we reviewed.

⁹ <http://dir.yahoo.com/>.

¹⁰ <http://www.dmoz.org/>.

¹¹ <http://www.taxonomywarehouse.com/>.

4 Implications for Revising WordNet's Domain Ontology

Our review of categorial organization in different resources provides a basis for re-thinking WordNet's current domain categories, their interrelations and possibilities for propagating the domain information throughout WordNet.

4.1 Selection of Domain Categories

We specify the following desiderata for WordNet domain categories:

1) Domain categories and members should be synsets. This assures, first, that the concept is unambiguously identified. Second, a domain category will automatically be linked to nodes accessible from its direct members via WordNet's relations. Ideally, a domain category should be a node with pointers to many other nodes.

2) Synsets that serve as domain categories as well as at least one of its members should occur frequently in text and speech, denoting a subject of learning or a common and familiar topic.

3) Domain categories should be chosen at the optimal level of generality. A category situated at the mid-level of a hierarchy typically is clearly distinctive from other categories on the same level and contains many sub-categories. Top level nodes in WordNet are too abstract and semantically unspecified; thus, {entity} would not be an informative domain category.

These properties could be translated into a set of rules that can guide the selection of domain categories. For example, if a candidate word has several senses, the one that is more frequent and more densely connected in WordNet is likely to be a better domain category label. For instance, {football, football_game} is to be preferred over a {football} (the ball).

A logical place for identifying domain label candidates is among frequent word forms in synsets found at the mid-level of WordNet's taxonomies.

4.1.1 Domains and Basic Level Categories. Rosch (1978) introduced the notion “basic category”¹², categories that occupy a special place in speakers’ conceptual organization as evidenced

¹² Unlike the Suggested Upper Merged Ontology (SUMO) <http://virtual.cvut.cz/kifb/en/toc/root.html>.

¹³ Basic categories are not to be confused with “Base Concepts” (BCs) in EuroWordNet (Vossen et al., 1998, Atserias et al., 2004). BCs are said to “play the most important role in the various WordNets of different languages” and characteristically occupy a high position in the semantic hierarchy and have many relations to other concepts.

in various cognitive and linguistic behaviors. Basic categories, located at the “basic” level in a taxonomy, are characterized by the fact that their members show high within-category similarity and low between-category similarity¹⁴. Basic categories thus have something in common with domain categories. Theoretically, basic-level categories could be identified in WordNet as those synsets with many sisters but few aunts (Izquierdo et al., 2007).

4.1.2 Automatically Extracting Domain Synsets. If we look at the domain taxonomies from the different resources summarized above, the division level tags are usually in line with the middle level domain ontology that is discussed in the previous subsection. This suggests that reoccurring domains across different resources are generally good seeds for the revision of WordNet’s domain ontology. As an experiment, we collected all the class-level and division-level tags from DDC, LCC, UDC, USAS, OED, Propædia and Taxonomy warehouse, and generated a list of 209 domain categories that can be mapped to WordNet synsets. This is only a small subset of domains that should be incorporated into the WordNet domain ontology. Additional methods should be applied to expand the list. However, there are great challenges in completely automating the domain selection process.

4.1.3 Prototypicality. Even if we were able to locate the levels in the WordNet hierarchy that meet the criteria of a basic-level category and qualify as domain categories, the automatic labeling of domain members raises some questions. Not all domain members are created equally. For example, {church, Christian church}, {Judaism}, {Hinduism}, {Taoism}, {Buddhism}, {Khalsa}, {Scientology}, {Shinto} are hyponyms of {religion}, but they are not universally equally salient: {Shinto} is a religion confined to Japan only and its membership in the {religion} domain seems intuitively weaker than that of {Hinduism}, for example. The prototype theory developed by Rosch (Rosch, 1978) provides a useful perspective on such cases. Although mainly tested on natural kinds (vegetables) and artifacts (furniture), Rosch’s work showed that speakers judge some members (prototypes) are better examples of their category than others. This suggests another criterion for domain status: in addition to frequency and number of outgoing poin-

ters, a mid-level synset in WordNet would thus make a good domain category if it is judged to be a prototypical member (i.e. cognitively salient) of its superordinate category. The same applies to members of a category. Some synsets like {e.g., edible_fruit} have many children, but not all are equally salient. Labeling of domain categories and members should thus be subject to human judgment after a candidate list is generated automatically.

4.1.4 Semantic Distance. In addition to prototypicality, synsets at the same level (sisters) may vary greatly in terms of their semantic distance to the upper level. For example, {established church}, {sect}, and {cult} are currently listed as sisters of specific religions like {Hinduism}, but they should be located on a separate, higher level. Second, human language does not lexicalize concepts in semantically equal intervals. Thus, the semantic distance between {possession} and three of its subordinates, {white elephant}, {treasure} and {property}, is intuitively very different; only {property} seems to qualify as a domain category. Thus, sisterhood in WordNet does not necessarily entail equal category status.

4.2 Construction of Domain Hierarchy

Although assigning synsets to domains may appear like a straightforward task, given WordNet’s hierarchies, it is not. For example, {football} and {baseball} intuitively are sisters; however, {baseball} is hyponym of {ball game}, a sister of {football}. In fact, when we mapped the 209 domain class and division labels from the different resources described in Section 3 to synsets in WordNet, we found that they spread across 10 levels -- a classification that is clearly too complex for a homogeneous domain ontology.

4.3 Propagation of Domain Category Membership throughout WordNet

The WNDomains project has proposed a way to transmit domain tags across WordNet network using existing relations. However, this process is restricted to members of a hierarchy that belong to the same part of speech, thus requiring the domain tags to be separately and manually seeded for nouns, verbs, and adjectives.

There are several possible approaches to assign domain member tags across different parts of speech in WordNet. Lexical overlap among the glosses and examples of WordNet synsets from different categories is a good indicator of semantic relatedness. For example, the synset {play: perform *music* on (a musical instrument)}

¹⁴ Fellbaum (1990) tried to characterize basic level in verb hierarchies, which are flatter than noun taxonomies and often show a clear “bulge” at the mid-level.

belongs to the domain *{music}*. Corpus-based co-occurrence data and topic models, which have been widely used for Word Sense Disambiguation, can further help to identify domain-related synsets from different parts of speech.

4.4 Evaluation of the Domain Ontology

WordNet's domain organization can be evaluated in several ways. One could map the domain ontology back to other domain category taxonomy resources and assess whether the WordNet domain information could provide a reasonable index for the different collections of knowledge.

Another method is to apply the domain information to various Natural Language Processing tasks, such as Word Sense Disambiguation, topic classification, text summarization, etc. If adding the WordNet-based domain tags improves the accuracy and speed of these tasks, a good argument for their meaningfulness and usefulness can be made.

5 Conclusions and Future Work

We reviewed the potentials and limitations of WordNet's current domain structure and WordNet Domains Project. Inspired by the domain taxonomies found in different resources, we propose general guidelines and methods for reconstructing a comprehensive and rich WordNet domain ontology. While WordNet's current relations can be harnessed for the identification and propagation of domain categories and their members, manual verification is indispensable. We discuss challenges and possible evaluation mechanisms of the proposed approach.

Acknowledgments

We thank the WordNet Domains Project for sharing their data.

References

- Archer, D., Rayson, P., Piao, S. and McEnery, A. M. 2004. Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies. In Proc. EURALEX-2004 conference. 817-827.
- Atserias J., Climent S., Rigau G. 2004. Towards the MEANING Top Ontology: Sources of Ontological Meaning. In Proc. LREC'04. Lisboa.
- Bentivogli, L., Forner, P., Magnini, B. and Pianta, E. Revising Wordnet Domains Hierarchy: Semantics, Coverage, and Balancing. In Proc. COLING 2004 Workshop on "Multilingual Linguistic Resources". 101-108. 2004.
- Fellbaum, C. 1990. English Verbs as a Semantic Net. *International Journal of Lexicography*. 3 (4): 278-301.
- Fellbaum, C. 1998. Wordnet: An Electronic Lexical Database. MIT Press. Cambridge, MA.
- Guthrie, J. A., Guthrie, L., Wilks, Y. and Aidinejad, H. 1991. Subject-Dependent Co-Occurrence and Word Sense Disambiguation. In Proc. the 29th annual meeting on Association for Computational Linguistics. 146-152.
- Izquierdo, R., Suárez, A. and Rigau, G. Exploring the Automatic Selection of Basic Level Concepts. In Proc. the International Conference on Recent Advances on Natural Language Processing (RANLP'07). 2007.
- Liddy, E. D., Paik, W. and Yu, E. S. 1993. Document Filtering Using Semantic Information from a Machine Readable Dictionary. In Proc. Workshop on Very Large Corpora, ACL1993. 20-29.
- Longman. 1978. *Longman Dictionary of Contemporary English*. Harlow, UK.
- Magnini, B. and Cavaglia, G. 2000. Integrating Subject Field Codes into Wordnet. In Proc. the 2nd Conference on Language Resources and Evaluation (LREC). 1413-1418.
- Miller, G. Nouns in Wordnet. 1998. In Fellbaum, C. (Ed.). *Wordnet: An Electronic Lexical Database*. MIT Press. Cambridge, MA.
- Miller, J. and Sears, M. E. 2004. *Sears List of Subject Headings*. H.W. Wilson Co. New York, NY.
- Rosch, E. Principles of Categorization. 1978. In Rosch, E. & Lloyd, B.B. (Ed.). *Cognition and Categorization*. Lawrence Erlbaum. Hillsdale, NJ.
- Siegel, E. V. Disambiguating Verbs with the Wordnet Category of the Direct Object. 1998. In Proc. The Usage of WordNet in Natural Language Processing Systems Workshop.
- Vossen, P., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., Roventini, A., Bertagna, F., Alonge, A. and Peters, W. 1998. *The EuroWordNet Base Concepts and Top Ontology*. Technical report, Paris, France.
- Walker, D. E. and Amsler, R. A. 1986. The Use of Machine Readable Dictionaries in Sublanguage Analysis. *Analyzing Language in Restricted Domain. Sublanguage Description and Processing*. Lawrence Earlbaum. Hillsdale, NJ.
- Wilks, Y. and Stevenson, M. 1998. Word Sense Disambiguation Using Optimised Combinations of Knowledge Sources. In Proc. the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, 2: 1398-1402.

An Implementation of a System of Verb Relations in *plWordNet* 2.0

Marek Maziarz

Institute of Informatics
Wrocław Univ. of Technology
marek.maziarz@pwr.wroc.pl

Maciej Piasecki

Institute of Informatics
Wrocław Univ. of Technology
maciej.piasecki@pwr.wroc.pl

Stan Szpakowicz

EECS, University of Ottawa &
ICS, Polish Academy of Sciences
szpak@eecs.uottawa.ca

Abstract

Verbs usually make up 10-15% of the material in a wordnet. We have embarked on a project in which this proportion increases to a lexically more realistic 1/3. The development of *plWordNet* 2.0 has been geared toward building up the verb hierarchy very considerably. The scaffolding upon which this part of our wordnet will rest is a detailed system of semantic relations. The selection of relations is informed by the linguistic accounts of Polish lexical semantics, notably the derivational procedures. The paper presents the background, rationale and specific solutions. While the system we describe has been tailored to one language, it could suggest similar solutions for typologically similar languages.

1 Introduction

1.1 Motivation

Verbs, the second-largest component of any wordnet, have properties which go far beyond lexical semantics, the principal concern in wordnets. Predicate-argument structure, semantic roles, subcategorisation, frames, inter-clause connections – such phenomena are what verbs are mainly about. That may be why verb relation systems vary across wordnets and versions, and appear less settled than inventories of relations among nouns.

The research question which we pose ourselves is this: how to steer a course between too much and too little verb semantics in a wordnet, and how to account for the lexical-semantic properties of verbs in an inflected language with highly productive derivation. It is an additional consideration that we work on a wordnet in which the basic building block is a lexical unit rather than a synset (Piasecki et al., 2009). Another factor is that aspectual differences pervade the verb system in our language, so aspect must be accounted for. All in all, we seek a set of relations for verbs which will ensure a clear account of all differences among lexical units – mandated by the language system and confirmed by links in a wordnet.

Princeton WordNet (PWN), a model for nearly all wordnets, has thus far paid much more attention to nouns than to verbs – and so, it seems, have other notable wordnets – see Table 1.¹ This is historically and practically justified. Many applications rely on the noun component and expect it to be highly developed. Yet verbs and common nouns may be more or less equally numerous in any language’s vocabulary. If a wordnet is to be a go-to resource for language processing, such relative neglect of verbs must end.

We have attempted to give verbs their due in the *plWordNet* 2.0 project. Not only does our wordnet already contain nearly 1/3 verbs (Table 1), but there also is in place a rich system of lexicosemantic relations among verbs, as well as from verbs to other categories. It is that system which will be the main theme of this paper.

1.2 Influences

Our endeavour began with a careful analysis of the relation sets in PWN, *EuroWordNet* (EWN) and *GermaNet*. We concluded soon enough that, while in theory comprehensive, none of those sets allows us distinctions fine enough to account for the intricacies of the Polish derivational system. This paper will first enumerate another 13 relations, a few of them with subtypes, and then discuss our decisions. The resulting system of relations should be broadly compatible with those sets and applicable, with few adjustments, to Slavic languages.

The theory of PWN (Fellbaum, 1998b, pp. 76-88, 220-223) lists six relations relevant to our investigation: synonymy, antonymy, inclusive entailment (proper inclusion), troponymy (coextensiveness), cause and presupposition.² In PWN 1.5 and later, inclusion and presupposition are combined into *entailment*.

GermaNet identifies troponymy with hyponymy (Kunze, 1999). PWN’s inclusive entailment is named *subevent* as in EWN (Alonge, 1996, p. 43),

¹Czech data cited after (Fišer, 2007).

²We omit *domain* and *group of verbs*.

VERBS	PWN 3.1 (June 2011)	<i>GermaNet</i> (Apr. 2011)	BulNet (Sep. 2005)	Czech WN	plWN 1.0 (fall 2009)	plWN 1.5 (Sep. 2011)
LUs	25047, 12.1%	12981, 13.9%	—	—	3497, 16.7%	31184, 32.0%
synsets	11529, 12.1%	9850, 14.2%	4421, 18.1%	5126, 18.0%	1860, 10.5%	17044, 24.7%

Table 1: Verb lexical units across selected wordnets.

and ‘entailment’ is only a label for backward presupposition (Hamp and Feldweg, 1997). Causality is cross-categorial (Kunze and Lemnitzer, 2002).

EWN’s set is similar to *GermaNet*’s (Vossen, 1998, p. 94). Hyponymy is PWN’s troponymy. Cause includes PWN’s presupposition (Vossen, 1998, p. 109). Synonymy is weak.³ EWN also introduced near-synonymy (for semantically close co-hyponyms), near-antonymy, as well as cross-categorial synonymy, antonymy and hypernymy.

There are 19 verb relations in *plWordNet*, 6 of them derivational – a set significantly built up from that in *plWordNet* 1.0. Table 2 contrasts our relations and relations shared with (or very similar to relations in) PWN, EWN and *GermaNet*. Details of the additions follow in Section 2.

2 Verb Relations in *plWordNet*

The core relations listed in Table 2 do not differentiate lexical units (LUs) sufficiently well. It is a fact confirmed by rather low network density of verbs in *plWordNet* 1.0 (Piasecki et al., 2010).⁴ We aim at increasing it significantly in *plWordNet* 2.0 by introducing a finer-grained system of relations.

We now list the new relations, those not taken over (one way or another) from PWN, EWN or *GermaNet*. Brief descriptions follow; in parentheses – the element linked by the relation (synset or LU) and parts of speech to which it applies. The fuller definitions, and more examples, appear in (Maziarz et al., 2011).

The purpose of the first six relations on our list is a further specification of the relations noted in wordnets. Except *converseness*, they are meant to enhance differentiation of synsets. *Converseness*, relatively frequent in the Polish data, was introduced to complete the description of antonymy.

- **Processuality** (synset, V-N, V-Adj) describes transition to a state described by a noun

³Two words are synonymous if they are interchangeable in at least one context (Vossen, 1998, p. 104).

⁴Network density is the average number of relation instances (links) between a LU and any other LU in the whole wordnet or, in our case, in its verb component.

or an adjective (zgłupieć ‘become_{pf} a fool’ or ‘become_{pf} stupid’ [głupiec ‘fool’, głupi ‘stupid’]).

- **Inchoativity** (synset, V-V) holds between a verb of ingressive/inchoative situation and a verb initiating that situation (zasnąć ‘fall_{pf} asleep’ – spać ‘sleep_{impf}’).
- **State** (synset, V-N, V-Adj) is best paraphrased as “X is Y_N” or “X is Z_{Adj}” (królować ‘rule_{impf} as a king’ = być królem ‘be_{impf} a king’).
- **Preceding** (synset, V-V) resembles presupposition but with a weaker form of precondition: typical rather than mandatory. Examples: siedzieć ‘sit_{impf}’ – stać ‘stand_{impf}’, siedzieć ‘sit_{impf}’ – leżeć ‘lie_{impf}’.
- **Inter-register synonymy** (synset, V-V) links verbs in significantly different registers (mówić ‘speak_{impf}’ – pieprzyć [vulgar] ‘talk_{impf} nonsense’).
- **Converseness** (LU, V-V) links converses – verbs with opposite meaning and mutually reversed roles of arguments (kupić ‘buy_{pf}’ – sprzedawać ‘sell_{pf}’).
- The remaining six relations are motivated by derivational facts in Polish.
- **Pure aspectuality** (LU, V-V) links pure derivational pairs (wykuć ‘forge_{pf}’ – wykuwać ‘forge_{impf}’).
- **Secondary aspectuality** (LU, V-V) links secondary derivational pairs (zgubić ‘lose_{pf}’ – gubić ‘lose_{impf} [habitual]’).
- **Cross-categorial synonymy** (LU, V-N, V-Adj) goes by derivation across parts of speech without change of meaning (mówić ‘talk_{impf}’ – mówienie ‘talking_{gerund}’ and mówiący ‘talking_{participle}’).

PWN	<i>GermaNet</i>	EWN	<i>plWordNet</i>
synonymy	synonymy	synonymy	synonymy (V-V)
antonymy	antonymy	antonymy	antonymy (V-V)
troponymy	hyponymy	hyponymy	hyponymy (V-V)
entailment (proper inclusion)	subevent	subevent	meronymy (V-V)
entailment (backward presupposition)	entailment	cause	presupposition (V-V)
cause	cause*	cause	causality* (V-V, V-N, V-Adj)
—	—	'near' relations	cross-categorial synonymy
—	—	fuzzynymy	fuzzynymy

Table 2: Core verb relations in selected wordnets (* = cross-categorial relations)

- **Multiplicativity** (LU, V-V), three subtypes:
Iterativity impf-impf links imperfective derivatives with imperfective bases (*jadać* ‘eat_{impf} from time to time’ - jeść ‘eat_{impf}’).
Iterativity impf-pf links imperfective derivatives with perfective bases (*zakochiwać się* ‘fall_{impf} in love occasionally’ - *zakochać się* ‘fall_{pf} in love’).
Distributivity links verb LUs with a clear distributive meaning with their bases (*pozabijać* ‘kill_{inf,pf} (many individuals)’ – *zabić* ‘kill_{inf,pf}’).
- **Role inclusion** (LU, V-N) links a denominational verb with its base, usually paraphrased as a thematic role – Agent, Patient, Instrument and so on (*pozłocić* ‘gild_{pf}’ < złoto ‘gold’, *posrebrzyć* ‘cover_{pf} with silver’ < srebro ‘silver’).⁵
- **Derivationality** (LU, V-V, V-Adj, V-N) is a last-resort, catch-all derivational relation.

3 Entailment

Troponymy in PWN, the inverse of verb hyponymy, is described as “To V1 is to V2 in some particular manner”, and called “a manner relation” in (Fellbaum, 1998a, pp. 79-80, 285, 213). Fellbaum states that troponymy links temporally co-extensive verbs; she argues that verb hyponymy differs from noun hyponymy, because noun tests cannot apply to verbs without essential changes.

EWN has opted for verb hyponymy instead. As in PWN, it must link verbs describing co-extensive

situations (Vossen, 2002, p. 35). The test expression “to X is to Y + AdvP/AdjP/NP/PP” (Vossen, 2002, p. 23) – “to *slurp* is to *eat* noisily” – resembles that in PWN, where ”AdvP/AdjP/NP/PP” is EWN’s rendition of “in some particular manner”. The manner component has its own representation in EWN structure. It links a hyponym with an adjective or adverb characterising it (*slurp* – *noisily*) (Vossen, 2002, p. 36). Vossen (2002, p. 13) claims that the term *hyponymy* may be successfully adapted to the manner relation.

Traditional semantics defines hyponymy and hypernymy via sense inclusion (Carter, 1998, p. 21), (Murphy, 2003, p. 236). In a modern formulation: hyponymy can be seen as predicate modification of a superordinate (*to swim is to move through fluid*) (Maienborn et al., 2011, p. 460). It seems that in PWN and EWN tests the hyponym is a *definiendum* and the hypernym is the head of a *definiens*). Verb hyponymy would thus be no different than noun hyponymy, even if test expressions for nouns and verbs are essentially different. It must be pointed out, however, that verb hierarchies are shallower than noun hierarchies (Carter, 1998, p. 21). This shows that verbal and nominal hyponymies do differ.

We assume, then, that – despite clear differences – verb hyponymy resembles noun hyponymy. As in EWN, we talk of *hyponymy* and *hypernymy* rather than entailment. The substitution test for verb hyponymy refers to infinitives and a manner element (AdjP/AdvP/NP/PP as in EWN and PWN):

X-ować to Y-ować w specjalny sposób
‘to X is to Y in a special way’

Thus the pairs *pływać* ‘swim’ – *przemieszczać się* ‘move’, *nawilzać* ‘moisten’ – *moczyć* ‘wet’,

⁵The relation has no inverse; *EuroWordNet*’s mutual inverses Role and Involved hold between synsets (Vossen, 2002, pp. 28-9).

gwałcić ‘rape’ – *krzywdzić* ‘hurt’ are in the hyponymy/hypernymy relation because of *pływać* = ‘move through water’, *nawilżać* = ‘wet slightly’ and *gwałcić* = ‘hurt by forcing to have sex’.

The denotation of a verb can be seen as a set of situations. Situations do not only have participants (roles) and a temporal span, but they often have constituents – sub-situations which together make up a more complex structure. Naturally, such constituents are denoted by other verbs. For example, *jeść* ‘eat’ has three sub-situation synsets: ⟨*gryźć* ‘chew_{impf}’⟩, ⟨*polkać, przelykać* ‘swallow_{impf}’⟩ and ⟨*połknąć, przełknąć* ‘swallow_{pf}’⟩. The connection between a situation as a whole and its constituents is captured by broadly conceived entailment. In *plWordNet* 2.0, we analysed other part-whole relations and decided to extend *meronymy* to verbs. We introduced two new subtypes of the meronymy, both defined only for verbs (Maziarz et al., 2011, p 191):

- «*Meronymy and holonymy of sub-situation* associate a composite situation and its component. There is “temporal inclusion” between the component and the whole.» The substitution test refers to the idea of sub-situation as an integral and typical part of the situation represented by the holonym.
- «*Meronymy and holonymy of accompanying situation* accounts for a “primary” situation, represented by the holonym, typically supplemented by another situation, represented by the meronym.»

Substitution tests for both subtypes also require that the situation described by the meronym be entailed by that of the holonym. “Temporal inclusion” is much simpler than what temporal logic allows, but it seems to be sufficient for distinguishing both subtypes. It is worth noting that verb meronymy shares with noun meronymy the basic idea of the association of the sub-components with the whole, but verb subtypes do not intersect with noun subtypes. Our goal was to recognise and name specific forms of entailments and identify their place in the high-level picture of our wordnet’s system of the lexico-semantic relations.

Verb meronymy in *plWordNet* corresponds to HAS_SUBEVENT in EWN, subevent in *GermaNet* and proper inclusion in PWN. EWN refers to subevent as a kind of entailment, compared to meronymy of “concrete entities” (Vossen, 2002, p.

MERONYMY	N	%
subsituational	179	7.5%
accompanying situation	34	1.4%
HOLONOMY	N	%
subsituational	2106	88.4%
accompanying situation	62	2.6%

Table 3: Percentage of mero-/holonymy subtypes in *plWordNet* 1.5

36). A key test is “*Y takes place during or as a part of X, and whenever Y takes place, X takes place*” (*ibid.*). It shows that the relation is of the temporal inclusion type (in PWN terms) and has part-whole inclinations. It must be undoubtedly identified with proper inclusion entailment of PWN: the time span of the subevent is included in the time span of superordinate situation (Vossen, 1998, p. 94). The ‘meronymic’ nature of proper inclusion is obvious (Fellbaum, 1998b, p. 78).⁶

Verb meronymy and holonymy together account for 9.4% of *plWordNet*’s verb synset relations. Holonymy of sub-situation is the most frequent, probably because it usually describes the manner of action (quite like the *manner* relation in EWN). For example, *chlipać* ‘lap_{impf} up [drink noisily, making smacking sounds]’ has a meronym *mlaskać* ‘smack_{impf} [make smacking noises]’ and is a hyponym of *pić* ‘drink_{impf}’. Co-hyponyms of *chlipać* include *cmoktać* ‘smack_{impf} [drink making lapping sounds]’ and *ssać* ‘suck_{impf}’. The former has a meronym *smakować* ‘relish _{impf} [food or drink]’, the latter – *wyciągać* ‘draw_{impf} out [e.g. liquid]’. The verbs *chlipać*, *cmoktać* and *ssać*, defined by hyponymy with *pić* (*genus proximum*), are distinguished only by their meronyms which express *differentia specifica*.

Presupposition, understood very strictly, can be attractive for potential applications, because it makes the precondition mandatory and true all over the contexts of use. The present version of *plWordNet* (September 2011) includes 136 presupposition links. Example: *dachować* ‘±capsize [applied to cars]’ – *jechać* ‘go [when applied to cars]’). The verb component is still under development, but the number of presupposition instances is unlikely to increase. In many cases, however, we en-

⁶*GermaNet* and EWN point out the differences between two types of PWN entailment (presuppositional and proper inclusion) on the basis of temporal inclusion and exclusion (Alonge, 1996, p. 43), (Hamp and Feldweg, 1997), (Vossen, 1998, p. 94), (Kunze, 1999).

counter a weak form of backward-going dependency, in which the precondition is desirable and typical, but not mandatory. Examples: ⟨gniewać się ‘be angry’⟩ – ⟨kłócić się ‘quarrel’, waśnić się ‘generate discord’, wadzić się ‘≈ generate discord’, sprzeczać się ‘argue’⟩. *Preceding* should be a much more frequent relation than presupposition. The potential problem is achieving consistency among wordnet editors.

Inchoativity works in the reverse direction: from a verb denoting an initial phase of situation *S* to a verb representing *S* as a lasting state or process. Example: ⟨nasuwać ‘draw [something upon something]’⟩ – ⟨pokrywać ‘cover’⟩. It seems moderately frequent (316 instances in *plWordNet 1.5*), but significantly different from *cause*: inchoativity instances do not express entailment.

4 Relations Motivated by Word Formation

The most frequent derivational association is *aspect*. We moved the discussion of its representation to a separate Section 5, because it is so central in the verb relation system. Derivational associations in Polish, quite like in other Slavic languages, are numerous and usually signal semantic distinctions. The *state* and *process* relations are also encoded by derivational associations, but both were expanded beyond the formal associations, and they now link non-derivative LUs and synsets. For example,

- *state* links *błekitnieć* 1 ‘be/appear_{impf} azure’ with its adjectival base *błekitny* ‘azure’;
- *state* links *czekać* ‘wait_{impf}’ and the adjective *gotowy* ‘ready’;
- *state* links *dyrygować* ‘conduct_{impf} [an orchestra]’ and its derivative *dyrygent* ‘conductor’;
- *state* links *lizusować* ‘toady_{impf}’ with its noun base *lizus* ‘toady’ and such synonyms as *pochlebca* ‘flatterer’ or *stużalec* ‘flunkey’;
- *process* links *błekitnieć* 2 ‘become_{impf} azure’ and the adjective *błekitny*;
- *process* links *wypływać* ‘become_{pf} famous [lit. surface]’ with such adjectives as *sławny* ‘famous’, *rozpoznawany* ‘recognized’;
- *owdowieć* ‘become_{pf} a widow or a widower’ has *process* links to *wdowa* ‘widow’ and *wdowiec* ‘widower’;

- *process* links ⟨*osłepnąć*, *ociemniećpf blind’ with ⟨*ślepy*, *niewidzący*, *niewidomy**

Role inclusion (1258 instances) is purely derivational. Instances: ⟨derivative, its base⟩, correspond to EWN’s *involved* relation and subtypes (Vossen, 1998, pp. 101-102). *Semantic role*, linking noun derivatives to the verb derivative bases, goes in the opposite direction but the relations are not mutual inverses. Both are directed and defined only for derivational pairs. Semantic roles affect the word formation process in Polish (Wróbel, 1998, 578-583) as in other languages (Duncan, 1985). Examples of verb-noun pairs (inclusion is very productive in Polish):

- *ocienić* ‘shade_{pf}’ < *cień* ‘shade’ (agent);
- *bębnić* ‘play_{impf} drums’ < *bęben* ‘drum’ (object);
- *cementować* ‘cement_{impf}’ < *cement* ‘cement’ (instrument);
- *pauzować* ‘pause_{impf}’ < *pauza* ‘pause, break’ (time);
- *plażować* ‘sunbathe_{impf}’ < *plaża* ‘beach’ (location)
- *faulować* ‘(commit a) foul_{impf}’ < *faul* ‘foul’ (result).

5 Aspect and Verb Classes

Aspect and semantic classes play a central role in our verb relation system. Word formation via aspectual derivation is listed among most prominent derivational relations in Slavic languages – see for example (Pala, 2008; Pala and Hlaváčková, 2007). Authors of Slavic wordnets tend to see aspect as a problem difficult to solve using wordnet relations (Pala, 2008; Pala and Smrž, 2004). Affixation carries not only aspectual differences but also lexical meaning shift (Raffaelli et al., 2008; Laskowski, 1998). The Czech WordNet deals with aspect in heterogeneous ways: it is captured by the HAS-SUBEVENT relation (Pala and Smrž, 2004), Pala also proposed special synset relations X HAS IMPF, X HAS PERF, X HAS ITER to help keep perfectives, imperfectives and iteratives in separate synsets. It is not obvious for every Slavic wordnet maker, however, whether aspectual

pairs should be kept in different synsets. In BulNet (Koeva, 2008) aspectual pairs were introduced in one synset at the beginning; Koeva claimed it to be insufficient in capturing lexical differences between many aspectual pairs, so a special morphosemantic *aspect relation* had to be added. As far as we know, no wordnet distinguishes pure and secondary aspectuality, although although the distinction has been noticed.

Pure and secondary aspectual pairs are distinguished in *plWordNet*. Aspectuality links morphologically connected pairs with the aspect values PERF and IMPERF. Pure aspectuality, in which the only semantic difference is the perfective-imperfective opposition, is interconnected with semantic classes, because after Laskowski (Laskowski, 1998) we consider pure aspectual pairs to be telic. In order to assist editors of *plWordNet*, rather sophisticated entailment-like tests for telicity and pure aspectuality have been developed, similar to those proposed by Vendler (1957)⁷, Comrie (1989) and Laskowski (1998). Here is the essential telicity test:

Jeżeli w ciągu jakiegoś czasu X-nął, to znaczy, że przez cały ten czas Y-ował

'If he X-ed for some time, then he must have been Y-ing for all that time'

Aspect affects verb relations. We impose the same aspect on synonyms, antonyms, converses, hyponyms and hypernyms. That is because verb pairs share a meaning element, usually the meaning of the superordinate verb (*genus proximum* of all sub-ordinates). For example, the pair *pogarszać* 'worsen_{impf}', *make_{impf} worse* > *zmieniać* 'change_{impf}' is a proper instance of hyponymy, but the hypernym *zmieniać* cannot be replaced by its aspectual counterpart *zmienić* 'change_{pf}': a perfective semantic element should not be included in an imperfective hyponymic verb.

One more example: the verbs *chorować* 'be_{impf} ill, suffer_{impf} from an illness' and *pochorować* 'be_{pf} ill [for some time]' have a lot in common: they both denote durative situation of being ill. They cannot, however, be put in the same synset because of the aspect difference. In fact, aspect signals an important semantic difference here – between an activity and a dynamic-changeable-atelic situation denoted by the prefix *po-*.⁸ A good

⁷Vendler's tests were analysed by Verkuyl (1989, p. 53).

⁸In Bulnet, *po*-formations are linked via hyponymy to their derivational bases (Koeva et al., 2010), which is unacceptable in *plWordNet*.

synonym for *chorować* is not *pochorować*. It is *cierpieć* 'suffer [from an illness]'. Like synonyms, antonyms must be hyponyms of the same hypernym, so they must have the same aspect. The verbs *kochać* 'love_{impf}' and *nienawidzić* 'hate_{impf}' are both hyponyms of *czuć* 'feel_{impf}'.

The verb classes are as follows:

- states (*imperfectiva tantum*, primarily atelic, static): *spać* 'sleep', *kosztować* 'cost',
- activities (*imperfectiva tantum*, primarily atelic, dynamic, changeless): *tańczyć* 'dance', *jeść* 'eat',
- accomplishments (*imperfectives and perfectives*: primarily telic), *budować* 'build_{impf}', - *wybudować* 'build_{pf}',
- achievements (*perfectiva tantum*, momentary verbs): *zgubić* 'lose (sth)', *klęknąć* 'kneel',
- dynamic-atelic-non-momentary perfectives (*perfectiva tantum*): *posiedzieć* 'sit (for a while)', *przesiedzieć* 'sit (for some time)', *nakłamać* 'lie (a lot)', *pozabijać* 'kill some number of people, animals'.

Our classification draws upon the post-Vendlerian typology (Vendler, 1957), notably on the work of Laskowski (1998)⁹ and Paducheva (1995). Although Vendler classified situations (usages) rather than verbs – see (Comrie, 1989, pp. 44-45) and (Verkuyl, 1996, p. 47) – we apply this typology to verbs. Our telicity test makes use of the prototypical semantic verb properties. For example, *budować* 'build_{impf}' and *wybudować* 'build_{pf}' are primary telic verbs, whereas the intransitive *tańczyć* 'dance_{impf}' and *zatańczyć* 'dance_{pf}' are primary atelic verbs. They can be made telic in sentences with bounded nominal arguments (objects), for example *zatańczyć walca* 'dance_{pf} a waltz' – we adopt Jackendoff's notion of *boundedness* (Verkuyl, 1996, pp. 230, 233). Semantic classes influence the distribution of relations: some apply to all verbs, others are limited to particular classes. Examples:

- Process links accomplishments with adjectives and nouns,¹⁰ for example, *zmętnieć* 'become_{impf} muddy [e.g., water]' and the perfective counterpart *zmętnić* are connected

⁹For the original class labels, see (Cetnarowska and Stawnicka, 2010)

¹⁰In rare cases also achievements and activities (iteratives).

with the adjective *mętny* ‘muddy [not transparent]’; *ramoleć* ‘become_{impf} a dotard’ and the perfective counterpart *zramoleć* are connected with *ramol* ‘dotard’.

- *Inchoativity* links inchoatives (achievements and activities) with states and activities: the achievement verb *nawrócić się* ‘convert_{pf} oneself’ is linked with the activity *wyznawać* ‘profess_{impf} [religion, ideology]’; the activity *nawracać się* ‘convert_{impf} oneself’ is linked with *wyznawać* by inchoativity.
- *Distributivity* (a type of multiplicativity) links distributive verbs (a sub-class of dynamic–atelic–non-momentary perfectives, usually created with *po-* prefixation) with perfective accomplishments or achievements: *pokraść* ‘steal_{pf} much’ and *ukraść* ‘steal_{pf}’.
- *Hyponymy* is limited not only to verbs of the same aspect, but also of the same semantic class. Let us reuse the preceding examples: *mętnieć* is a hyponym of *stać się* ‘become_{pf}’ – both are accomplishments; the activity verb *nawracać się* is a sub-ordinate verb of the activity hypernym *zmieniać się* ‘change_{impf} oneself’ (an iterative meaning).

Aspect and verb classes help form *plWordNet*’s structure, but the restrictions on relations are sometimes difficult to follow, e.g., perfective verbs with a delimitative *po-* are often defined in (Dubisz, 2004) by imperfectives: *ponarzekać_{pf}* ‘complain_{impf} [for some time]’ – *narzekać* ‘complain_{impf}’; *polatać_{pf}* ‘fly_{impf} [for some time]’ – *latać* ‘fly_{impf}’. It is not uncommon, though, for such verbs to get a proper definition using a perfective hypernym: *spędzić* ‘spend_{pf} [some time]’ or *pobyć* ‘stay_{pf} [for some time]’; *plWordNet* editors must then choose the right (perfective) hypernym for all *po-*formations even though Polish dictionaries are not aware of it. Also verb classes are often neglected in dictionaries. The accomplishment *przenieść* ‘carry_{pf} [from-to]’ is defined in (Dubisz, 2004) by an achievement: *zabrać* ‘take_{pf}’, while the achievement *spaść* ‘fall down_{pf}’ is explained with an accomplishment: *przemieścić się* ‘move_{pf} [from-to]’.

6 Conclusions

Verb description in wordnets tends to be limited in terms of network density of verb relations and

the number of lexical units. That is not the case for *plWordNet*. Our rich system of verb relations merges selectively and expands several existing wordnet systems of verb relations. We have kept the main semantic distinctions which underpin verb relations in other wordnets, and introduced relations which better differentiate verb LUs. We explored two sources of such new relations: a more detailed treatment of the general entailment relation, and semantic oppositions encoded by derivational relations. The two sources overlap, because several new entailment-like relations are supported by regular derivational associations. We extend meronymy and hyponymy to verbs, and – following EWN’s tradition of a rich set of cross-categorial relations – we continue to build a more unified wordnet model. Aspect and verb semantic classes help improve wordnet structure, and promote consistency among wordnet editors. This practice makes *plWordNet* unique among wordnets.

Acknowledgments

Financed by the European Union within European Innovative Economy Programme project POIG.01.01.02-14-013/09

References

- Atonietta Alonge. 1996. Definition of the links and subsets for verbs, EuroWordNet Project LE4003, Deliverable D006. (unpublished).
- Roland Carter. 1998. *Vocabulary. Applied Linguistic Perspectives*. Routledge, 2nd edition.
- Bożena Cetnarowska and Jadwiga Stawnicka. 2010. The verb’s semantics and its compatibility with temporal durative adverbials in polish. *Studies in Polish Linguistics*, 5:27–50.
- Bernard Comrie. 1989. *Aspect. An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge University Press.
- Stanisław Dubisz, editor. 2004. *Uniwersalny słownik języka polskiego* [a universal dictionary of Polish], electronic version 1.0. PWN.
- Jill Carrier Duncan. 1985. Linking of thematic roles in derivational word formation. *Linguistic Inquiry*, 16(1).
- Christiane Fellbaum. 1998a. A Semantic Network of English: The Mother of All WordNets. *Computers and the Humanities*, 32:209–220.
- Christiane Fellbaum, editor. 1998b. *WordNet – An Electronic Lexical Database*. The MIT Press.

- Darja Fišer. 2007. A multilingual approach to building slovene wordnet. In *Proceedings of the workshop on A Common Natural Language Processing Paradigm for Balkan Languages held within the Recent Advances in Natural Language Processing Conference RANLP'07*, Borovets, Bulgaria.
- Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a Lexical-Semantic Net for German. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Madrid.
- Svetla Koeva, Svetlozara Leseva, Ekaterina Tarponanova, Borislav Rizov, Tsvetana Dimitrova, and Hristina Kukova. 2010. Bulgarian Sense-Annotated Corpus – Results and Achievements. In *The Seventh International Conference. Formal Approaches to South Slavic and Balkan Languages 4-6 October 2010*, Dubrovnik, Croatia.
- Svetla Koeva. 2008. Derivational and Morphosemantic Relations in Bulgarian Wordnet. In *Proc. Intelligent Information Systems 2008*, page 359–368.
- Claudia Kunze and Lothar Lemnitzer. 2002. GermaNet – representation, visualization, application. In *Proc. LREC 2002, main conference*, volume V, pages 1485–1491.
- C. Kunze. 1999. Semantics of verbs within GermaNet and EuroWordNet. In E. Kordoni, editor, *Workshop Proceedings of the 11th European Summer School in Logic, Language and Information*, pages 189–200.
- Roman Laskowski. 1998. Kategorie morfologiczne języka polskiego – charakterystyka funkcjonalna [The morphological categories of Polish – the functional characterisation]. In Renata Grzegorczykowa, Laskowski Roman, and Henryk Wróbel, editors, *Gramatyka współczesnego języka polskiego. Morfologia* [The Grammar of Contemporary Polish. Morphology], volume 1. PWN, 2nd edition.
- Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors. 2011. *Semantics. An International Handbook of Natural Language Meaning*, volume 1. Walter de Gruyter GmbH.
- Marek Maziarz, Maciej Piasecki, Stanisław Szpakowicz, Joanna Rabiega-Wiśniewska, and Bożena Hojka. 2011. Semantic Relations Between Verbs in Polish Wordnet 2.0. *Cognitive Studies*, 11:183–200. www.eecs.uottawa.ca/~szpak/pub/Maziarz_et_al_CS2011b.pdf.
- M. L. Murphy. 2003. *Semantic Relations and the Lexicon*. Cambridge University Press.
- Elena V. Paducheva. 1995. Taxonomic categories and semantics of aspectual opposition. In Pier Marco Bertinetto, editor, *Temporal reference, aspect and actionality*, volume 1, pages 71–89. Rosenberg & Sellier.
- Karel Pala and Dana Hlaváčková. 2007. Derivational Relations in Czech WordNet. In *Proceedings of the Workshop on Balto-Slavonic Languages*, page 75–81, Prague.
- Karel Pala and Pavel Smrž. 2004. Building Czech Wordnet. *Romanian Journal of Information. Science and Technology*, 7(1-2):79–88.
- Karel Pala. 2008. Derivational Relations In Slavonic Languages. In *The Sixth International Conference: Formal Approaches to South Slavic and Balkan Languages, 25-28 September 2008*, Dubrovnik, Croatia.
- Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wrocław University of Technology Press. www.site.uottawa.ca/~szpak/pub/A_Wordnet_from_the_Ground_Up.zip.
- Maciej Piasecki, Stan Szpakowicz, and Bartosz Broda. 2010. Toward plWordNet 2.0. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors, *Principles, Construction and Application of Multilingual Wordnets. Proceedings of the 5th Global Wordnet Conference*, pages 263–270, Mumbai. Narosa Publishing House.
- Ida Raffaelli, Marko Tadić, Božo Bekavac, and Željko Agić. 2008. Building the croatian wordnet. In *Proceedings of the Fourth Global WordNet Conference*, Szeged. Global WordNet Association.
- Zeno Vendler. 1957. Verbs and times. *The Philosophical Review*, 66(2):143–160, April.
- Henk J. Verkuyl. 1989. Aspectual classes and aspectual composition. *Linguistics and Philosophy*, 12(1):39–94.
- Henk J. Verkuyl. 1996. *A Theory of Aspectuality. The Interaction between Temporal and Atemporal Structure*. Cambridge University Press.
- P. Vossen, editor. 1998. *EuroWordNet. A multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.
- Piek Vossen. 2002. EuroWordNet General Document Version 3. Technical report, Univ. of Amsterdam.
- Henryk Wróbel. 1998. Czasownik [The Verb]. In Renata Grzegorczykowa, Laskowski Roman, and Henryk Wróbel, editors, *Gramatyka współczesnego języka polskiego. Morfologia* [The Grammar of Contemporary Polish. Morphology], volume 2. PWN, 2nd edition.

Approaching *plWordNet* 2.0

Marek Maziarz

Institute of Informatics
Wrocław Univ. of Technology
marek.maziarz@pwr.wroc.pl

Maciej Piasecki

Institute of Informatics
Wrocław Univ. of Technology
maciej.piasecki@pwr.wroc.pl

Stan Szpakowicz

EECS, University of Ottawa &
ICS, Polish Academy of Sciences
szpak@eecs.uottawa.ca

Abstract

The Polish Wordnet, *plWordNet*, has been in steady development for five years. We are building it from scratch, all the time making provisions for its general compatibility with the other major wordnets. We are very close to reaching a milestone of 100000 lexical units in 70000 synsets. In addition to a fairly comprehensive coverage of common nouns, there already is in *plWordNet* a significantly built-up verb component, and a similarly enlarged adjective component is under construction. We present the background, the assumptions, the relation set (essential for any wordnet, and central for our inflection- and derivation-rich language) and the current state of the project, and we map the near future.

1 Promises

In 2009, we began the work on *plWordNet* 2.0, the next major release of the first large, publicly available Polish wordnet. The construction of *plWordNet*, initiated in 2005, has led to the release of a wordnet with 26990 lexical units in 17695 synsets (Piasecki et al., 2009, Section 5.2); *plWordNet* 2.0 has been planned (Piasecki et al., 2010b) as a very significant step toward a wide-coverage wordnet for Polish. We envisaged the expansion of *plWordNet* 1.0 in size and in the expressive power of the relation-based description. A special focus was to be given to verbs and adjectives, under-represented in the 2009 release.

While no one can venture a guess at the ideal size of a wordnet, an optimistic target is to exceed the size of the largest existing dictionaries: a wordnet should describe lexical units which occur in textual data relevant to its numerous applications. The size of the *Princeton WordNet* (PWN) is still a hard-to-reach target for other wordnets. Our initial plans for *plWordNet* 2.0 were to make it comparable in size to what were then large European wordnets, among them *GermaNet* (Kunze and Lemnitzer, 2002), and to include all most frequent Polish lemmas. That meant 70000-80000 lexical units

(LUs)¹ in 45000-55000 synsets. We revised our objective after receiving additional funds for the expansion of *plWordNet*: ≈ 135000 LUs in 90000-100000 synsets. We are already nearing 100000 LUs and 70000 synsets. The verb component is ready, and so is most of the noun component. Section 4 shows the detailed statistics.

The inventory of relations for nouns in *plWordNet* has been slightly modified; major changes are in place for verbs, and will be implemented for adjectives. Section 2 discusses the main assumptions and principles upon which *plWordNet* is founded. Section 3 briefly presents the system of relations. Section 4 presents the construction process. Finally, we discuss the experience gained and the work schedule for the last phase of the expansion.

2 Assumptions and Principles

PWN and most of other wordnets are structured into synsets. The synset is usually briefly described as capturing a lexicalised concept. A synset should contain a group of near-synonyms and represent the concept behind them, so that synset members share *some* meaning. How much is to be shared is left to the discretion of wordnet editors. An operational definition of synonymy is hard to formulate in a way which would support consistency of decisions among synset authors.

Synsets are linked by *conceptual relations* with names borrowed from linguistic work on lexical semantics, such as *hypernymy* or *meronymy*. Many lexico-semantic relations, however, clearly link LUs rather than sets of LUs (examples include various oppositions – including antonymy – and forms of derivation), and most wordnets note such links.² In fact, lexical semantics tends to say that hypernymy, meronymy etc. link pairs of LUs, not pairs of sets of LUs. We found it difficult to define simultaneously synonymy, synset and conceptual

¹Without going into details, a lexical unit can be understood as a lemma with a sense number.

²See *lexical relations* (Fellbaum, 1998, p. 17) or *relations between wordforms* (Miller et al., 1990), contrasted with *conceptual relations* or *relations between word meanings*.

relations. We proposed to adopt the LU rather than the synset as the centrepiece of the wordnet structure (Derwojedowa et al., 2008; Piasecki et al., 2009). Thus, lexico-semantic relations between LUs are primary, and from them we derive relations between synsets. LUs in a synset share certain (carefully selected) lexico-semantic relations: recognised by semanticists, well grounded in the wordnet tradition, frequent in language, with a reasonable *sharing factor*,³ and with a potential to facilitate wordnet applications. They come with linguistically accurate *substitution tests* (Vossen, 2002; Piasecki et al., 2009), so a group of editors can annotate them consistently.

Synsets in *plWordNet*, then, are a notational convenience: to say that synsets S_1 and S_2 are linked by relation R is to say that any pair $s_1 \in S_1$ and $s_2 \in S_2$ is an instance of R . We refer as *constitutive relations* to the lexico-semantic relations selected to be the basis of synset construction. Different parts of speech require different sets of constitutive relations; see Section 3. Our experience has also shown convincingly that additional criteria are necessary to distinguish precisely between members of any two synsets. Such secondary factors include stylistic register for nouns, and semantic class and aspect for verbs (Maziarz et al., 2011a; Maziarz et al., 2011b).

The *plWordNet* project has always focussed on lexico-semantic facts specific to Polish. We decided to forgo the route which many wordnet developers take: translate PWN and adjust the result of that translation. Not only are we building the whole network (Piasecki et al., 2009), but we also design from scratch a system of relations to underpin *plWordNet*. Register and aspect (richly manifested in Polish) are among the prime consideration, though we constantly keep in mind the future – inevitable – alignment with PWN.

The hypernymy structure in PWN was initially a forest with *unique beginners*, only later joined into a tree. Potential links among very general LUs (such as *entity* or *abstraction*) are seldom well motivated by linguistic criteria. Not all abstract notions are lexicalised, so the introduction of *artificial lexical units* may be required. That is why in *plWordNet* we only introduce those hypernymy links which are compatible with the lin-

guistic definition of hypernymy, and for which LU pairs pass the relevant substitution tests. This strategy must result in a hypernymy forest. On the other hand, *plWordNet* applications can only benefit from a single-root tree organisation (wordnet-based word-similarity calculation is a case in point). To meet application needs, we plan to map the top synsets (not linked by hypernymy or any other constitutive relation) onto a general, top-level ontology. SUMO (Niles and Pease, 2001) is among those considered.

We focus on the description of the Polish lexical system, so proper names (PNs) get very limited coverage in *plWordNet*. PNs are a very large, open category, which changes dynamically. Even if we wanted to select a limited subset, reasonably based on high frequency in a very large corpus, selection would be strongly biased by the origin of texts. We also wanted first to achieve nearly comprehensive coverage of “feasibly numerous” categories, mainly common nouns and verbs. Besides, the techniques of Named Entity Recognition support PNs and provide their classification. We made one exception: PNs which represent geographical objects and areas, and are the derivative bases for common nouns (such as inhabitants, for example “warszawianin” *Warsaw citizen* from “Warszawa” *Warsaw*) – a process very productive in Polish. Geographical names are a necessary completion of the description of the derivative common nouns.

In the relation-based paradigm of the lexical-semantic description, the number of relation links associated with a LU characterises well the amount of information encoded for this LU – each link adds to its differentiation from other LUs. In order to get good and balanced coverage of the description of LUs, we should aim at a wordnet which has at least several links for any LU. Piasecki et al. (2010b) propose to characterise this property by *network density*: the average number of relation instances – links – going from a LU to any other LU in the wordnet. Network density can be increased simply by increasing the number of relations, but an excessively detailed relation list would lead to an excessively fragmented description. The properties we postulate for the constitutive relations (Section 2) seem to be a good basis for selecting lexico-semantic relations for a wordnet. One constraint must be relaxed: we should not expect all wordnet relations to have high values of the sharing factor. For example, antonymy

³The sharing factor of a relation is the average size of a group of LUs which share this relation. Thus antonymy’s sharing factor is 1 (a LU has at most one antonym) and hypernymy’s usually well above 1.

is quite frequent but does not form LU groups.

At the early stages of the *plWordNet* 1.0 project there was no electronic dictionary on which we could base our work, so we adopted a corpus-based approach. The following recapitulation sums up a longer discussion in (Piasecki et al., 2009). First, lemma frequencies are generated from a very large corpus,⁴ previously analysed morpho-syntactically, lemmatised and disambiguated. Next, proper nouns are filtered out, using a few large gazetteers (Marciniuk and Piasecki, 2011). A morphological guesser is applied during the morpho-syntactic processing, so the list can include potential lemmas absent from the existing dictionaries. Lemmas with the highest frequency are selected if they are not yet in *plWordNet*. We usually take ≈ 9000 new lemmas in each phase of *plWordNet* expansion.

The corpus-based procedure allows us to include contemporary lexical units in the wordnet. In practice, however, every corpus is somewhat unbalanced, and that introduces a bias into the lemma frequency lists. That is why the process now includes consultation with dictionaries to correct flaws in corpus-derived frequency lists, though lemmas extracted from the corpus dominate. The reliance on the corpus imposes bottom-up direction on the construction of the wordnet hypernymy structure. There is no predefined hypernymy structure to import. Instead, LUs created for lemmas (extracted from the corpus) trigger the addition of synsets to link to the already existing hypernym synsets or to those recently added.

It is efficient, if linguistically not quite proper, to import language data from monolingual and bilingual dictionaries and existing wordnets. We could not, clearly, rely only on lemma frequencies and simple concordance. In a semi-automatic approach to wordnet development, we implemented several methods of extracting from our very large corpus potential instances of lexico-semantic relations. These raw data, combined in the *WordnetWeaver* system (Piasecki et al., 2009; Piasecki et al., 2011), suggest, for each new lemma, one or more LUs (Section 4). *WordnetWeaver* has been the main software support for the work of the editors, who still make all editing decisions but in a more efficient manner. Let us only note here that such partial automation complements

⁴The present version contains 1.2 billion tokens taken from several publicly available Polish corpora – see (Piasecki et al., 2009) – plus texts collected from the Internet.

the corpus-based development philosophy: automated tools ensure advanced semantic browsing and exploration of the language data and produce a condensed description of the discovered lexico-semantic dependencies for the editors.

3 Relations

The system of relations in *plWordNet* has been fundamentally informed by the solutions in PWN and *EuroWordNet* (EWN), but also substantially influenced by the Polish linguistic tradition and the assumption that the lexical unit is the basic building block. Nine central relations have been defined for *plWordNet* 1.0 (Piasecki et al., 2009), not counting synonymy implicitly encoded in synsets; with subtypes of meronymy and holonymy the actual number was 19. Network density was relatively high for nouns, but too low for verbs and adjectives. There were also clearly fewer verb relations than in PWN and EWN. The rich Polish derivation was given in *plWordNet* 1.0 only two very general relations: *relatedness* and *pertainymy*. The *plWordNet* relation system is now much more involved: 17 relations among synsets and 16 among LUs, plus synonymy. With subtypes, there now are 44 synset and 42 LU relations. Many of them have a derivational character or originate from the derivational relations.

3.1 Synset relations

Synset relations are lexico-semantic relations extrapolated from the level of LU via the sharing of a relation between candidate synset members (Section 2). Substitution tests have been defined for each relation and relation subtype.

Hypernymy/hyponymy is defined for all parts of speech, only for LUs (extrapolated onto synsets) of the same part of speech. For nouns, the relation's definition is very similarly to that in *EuroWordNet*, see (Maziarz et al., 2011a). A handful of hypernymy/hyponymy instances in the adjective component of *plWordNet* 1.0 have yet to be revised. For verbs, we have decided to follow the practice of *plWordNet* 1.0, inspired by *EuroWordNet*, and refer as verb hypernymy/hyponymy to a special kind of entailment. The test was enriched with constraints which force both linked LUs to have the same aspect and belong to the same semantic verb class, see (Maziarz et al., 2011b).

Inter-register synonymy (defined between nouns and between verbs, considered for adjec-

tives) is «synonymy between lexical units which have different stylistic registers» (Maziarz et al., 2011a). It is used to link stylistically marked lexical units with their unmarked counterparts.

Holonymy/meronymy (nouns and verbs) is divided into subtypes. We have kept *part*, *place*, *portion*, *element of a collection* and *substance*, defined in *plWordNet 1.0*. A new subtype, *taxonomic unit*, expresses «lexico-semantic relations inside scientific taxonomies, especially biological taxonomy, for example, *kotowane ‘felidae’* – *kotokształtne ‘feliformia’*» (Maziarz et al., 2011a).

By analogy, holonymy/meronymy has been adopted for verbs. Two subtypes, *accompanying situation* and *sub-situation*, link verb LUs. The first «accounts for a ‘primary’ situation, represented by the holonym, typically supplemented by another situation, represented by the meronym» (Maziarz et al., 2011b). The second «associates a composite situation and its component», referring to a kind of temporal inclusion between the component and the whole (*ibid.*); it corresponds to the *subevent* relation in *GermaNet* (Kunze, 1999) and EWN (Vossen, 2002). For example, the verbs *trząść* ‘shake [while travelling in a vehicle]’ and *jechać* ‘travel [in a vehicle]’ are linked by *accompanying situation*, while *kryć* ‘seek [in the hide-and-seek game]’ and *bawić się w chowanego* ‘play hide-and-seek’ are connected by *sub-situation*. The two differ in that a typical situation of travelling in a vehicle *need not* be accompanied by shaking, whereas seeking *is* a typical (obligatory!) part of the hide-and-seek game.

Type/instance links synsets made up of proper names (synonymous names put into the same synset) to nouns which are their most specific descriptions. For example, ⟨*Wrocławmiasto ‘city’*⟩ are linked by *type* relation. This is how it is done in *WordNet* (Miller and Hristea, 2006) and *EuroWordNet* (Vossen, 2002), where it is the relation *belongs_to_class / has_instance*.

Inhabitant (for nouns) arises from a specific but surprisingly productive derivational relation. Examples: *domownik* ‘household member’ – *dom* ‘house’, *wrocławianin* ‘one living in Wrocław’ – *Wrocław*. Because of the proper name variants and synonymous proper names, the relation was expanded beyond the derivational associations to link synsets, required to include the derivative and its base, respectively (Maziarz et al., 2011a).

The remaining relations, meant for verbs, are

described in detail by Maziarz et al. (2011b). For full definitions and motivation please refer to that paper, from which we also took the «» quotations.

Cause (from verbs to verbs, nouns, adjectives) is a form of entailment, signalled in dictionary descriptions by verbs synonymous to “cause”. The relation resembles *cause* relations in PWN (Fellbaum, 1998) and EWN (Vossen, 2002). There are two subtypes of *cause* for verb-to-verb pairs (pf-to-pf and impf-to-impf); four *cause of process* subtypes link verbs of different aspects with nouns and adjectives (pf-to-Adj, impf-to-Adj, pf-to-N, impf-to-N); and the *cause of state* subtype links perfective with imperfective verbs denoting states (pf-to-state), for example *usnąć ‘put to sleep_{perf.}* – *spać ‘sleep_{imperf.}*’. This variety of subtypes, a little paradoxically, helps maintain coherence between editors: it simplifies test expressions.

Process (from verbs to nouns or adjectives) associates «verbs which denote spontaneous change of state or any dynamic situation» with nouns and adjectives describing the result of the change. The relation can be paraphrased using the verb *become*. It links synsets, but is often indicated by derivational associations, e.g., it links the synsets ⟨*chamieć* ≈ ‘become_{imperf.} a boor’⟩ and ⟨*prostak* ‘simpleton’, *cham* ‘boor’, *wieśniak* ‘yokel’⟩. Four subtypes are defined by two values of aspect and two parts of speech (nouns and adjectives).

Inchoativity links verbs which describe either entering into a state or beginning an activity with verbs which describe *being* in this state or activity (in general, dynamic durative situations). There are two subtypes: perfective → imperfective and imperfective → imperfective verbs. An example is a link from ⟨*usypiać ‘put to sleep_{impf.}*, *zasypiać ‘fall asleep_{impf.}*⟩ to ⟨*spać ‘sleep_{impf.}*⟩.

State (from verbs to nouns or adjectives) expresses *being in a state*. It links stative verbs (representing static situations) with nouns or adjectives which describe a state. An example: *panować ‘rule_{inf.}*’ is to be *pan ‘lord, ruler’*.

Multiplicativity is a relation of a derivational character, with subtypes. *Iterativity impf-impf* «can link pairs of imperfective verbs such that one of them, which expresses an iterative meaning, is derived by suffixation from the other», for example, ⟨*pisywać 1 ‘write_{impf.} sometimes’*⟩ and ⟨*pisać 1 ‘write_{impf.}*⟩. *Iterativity impf-pf* subtype «can also link imperfective derivatives of *perfectiva tantum*»; for example, *zakochywać się*

‘fall_{impf} in love sometimes’ is the iterative form of *zakochać się* ‘fall_{pf} in love’. The third subtype, *distributivity*, associates a perfective verb which represents multiplicative performance of an action on many patients or by many agents with a perfective verb which denotes the performance of the whole process; for example, ⟨*nałożyć* ‘catch_{pf} (plenty of)’⟩ and ⟨*złowić* ‘catch_{pf}’⟩.

Presupposition «expresses the backward-going dependency between a situation represented by the given verb and a situation whose occurrence is a kind of precondition». The precondition is «mandatory regardless of the negative polarity of the sentence with the given verb». Example: ⟨*dawać* ‘give’⟩ and ⟨*mieć* ‘have’⟩.

Preceding is similar to *presupposition*, but the precondition is treated as desirable or holding in many, but not necessarily all, situations. Example: ⟨*popuścić* 1 ‘loosen’⟩ and ⟨*ścisnąć* 1 ‘press (together)’, *zacisnąć* 1 ‘tighten’⟩.

All synset relations except inter-register synonymy are treated in *plWordNet* as constitutive: they meet the conditions defined in Section 2.

3.2 Lexical unit relations

Lexico-semantic relations not extrapolated to the level of synsets comprise two large groups: relations (motivated by linguistic or wordnet traditions) which do not express a sharing factor, and derivational relations. Synonymy is not directly described but it is encoded by synsets.

Antonymy applies to all parts of speech. It has been defined similarly to the definitions in PWN and EWN (Maziarz et al., 2011a), but divided into two subtypes: *complementarity* and *gradable opposition*. Complementary antonymy includes bipolar pairs of LUs with opposite and exclusive meanings, for example *man* – *woman*. Gradable antonymy links LUs with opposite senses which do not exhaust the semantic field, for example, *abstinent* 1 ‘teetotaller’ – *pijak* ‘drunkard’.

Converseness is a relation of oppositeness, applicable to nouns and verbs (Cruse, 1986, 10.6–10.7). It was considered in (Fellbaum, 1998), but in the end not included in PWN. For verbs, it is signalled by the mutually opposite roles assigned to the arguments, as in the classic pair *sell* – *buy*. Nouns are converses if they play opposite roles in some situation. A good substitution test is “If A is X (Prep) B, then B is Y (Prep) A” where X and Y are the nouns under investigation, such as X=*wife*

and Y=*husband*.

Cross-categorial synonymy, always expressed by productive derivational patterns, has been defined for the noun-verb, noun-adjective and verb-adjective pairings. It has six subtypes, because the relationship is directional.

The **feature bearer** relation links a noun which represents an object characterised by some feature to an adjective which represents the feature, for example *starzec* ‘an old man’ – *stary* ‘old’. **State** is an inverse of *feature bearer*.

Femininity links a feminine noun LU to its masculine derivational base. This relation, exemplified in English by *actress* – *actor*, is quite productive in Polish.

Markedness captures several forms of emotional markedness in derivationally associated nouns. There are three subtypes, all linking marked to unmarked words: *diminutive* (*armatka* ‘small cannon’ – *armata* ‘cannon’), *augmentative* (*brzuch* ‘big belly’ – *brzuch* ‘abdomen’), and *young being* (*wilczek* ‘wolf cub’ – *wilk* ‘wolf’).

Semantic role (noun to verb, as well as noun to noun) «characterises associations between a noun and derivationally linked verb from the perspective of a situation denoted by the verb» (Maziarz et al., 2011a). This relation is very similar to the *role* relation in EWN, and similarly subdivided into *agent*, *patient*, *instrument*, *location*, *product*, *time*, *agent of hidden predicate*, *object (of hidden predicate)* and *product (of hidden predicate)*. The last three subtypes are defined for derivationally associated pairs of nouns. The relation is directional: from a derivative to its base.

Role inclusion (verb to noun) is semantically opposite to **semantic role**, but the two relations are not mutually inverse. That is because they are always defined only for pairs: derivative and its base. The *role inclusion* subtypes are analogous to the first six subtypes of **semantic role**.

The **derivational** relation and **fuzzynymy** apply to all parts of speech. As in EWN, they are the last resort: the editor is convinced that two LUs are somehow related, but no regular *plWordNet* relation “works”. Fuzzynymy is extrapolated to synsets by the relation-sharing rule.

4 The Construction Process

The growth of *plWordNet* from 27000 LUs to nearly 100000 LUs required the average workload of about 3.5 full-time editor positions over 1.5

years. It is notable that we did not resort to any form of translation from another wordnet, nor to importing data from any lexico-semantic resource. The construction process relied on the processing of a very large corpus. The editors could consult several dictionaries (Dubisz, 2004; Bańko, 2000) when they edited suggestions generated by the automated tools. We are confident that the fast pace of work was greatly assisted by the organisation of work we adopted, and by significant software support for the work of the linguists.

The work was divided into phases of 3-5 months. Each phase concentrated on the part of the network for one category (noun, verb, adjective). The first step was to firm up the definition of the relation system for this part of speech. This inevitably had wider consequences: there are many cross-categorial relations, so any change may affect relations for other parts of speech. Next, substitution tests are required for each relation and its subtypes. Tests – with a very strict structure – are treated as an intrinsic part of relation definitions. They are automatically instantiated with specific lemmas for testing, and systematically presented to the editor in a wordnet-editing system called *WordnetLoom* (Piasecki et al., 2009; Piasecki et al., 2010a). The number of relation subtypes had increased considerably because of the need to make test specifications formal.

Next, we select lemmas for addition to the wordnet and prepare knowledge sources which describe those lemmas for the automatic tools. Lemmas are extracted from our corpus.⁴ During the first phases of *plWordNet* expansion, lemmas not recognised by the morphological analyser were filtered out; later we left on the list very frequent lemmas recognised by the morphological guesser. Next, we prune all proper names found in a large gazetteer (Marcińczuk and Piasecki, 2011). We always select 7000-9000 most frequent lemmas.

For the selected lemmas – combined with the lemmas MP already included in *plWordNet* – the following information is automatically produced from the corpus (Piasecki et al., 2009):

- Measure of Semantic Relatedness (MSR),
- lemma pairs extracted by hand-written lexico-syntactic patterns designed to detect hypernymy,
- lemma pairs extracted by automatically discovered statistical lexico-syntactic patterns,

- a classifier (trained on the data extracted from the corpus) designed to distinguish instances of *plWordNet* relation and other lemma pairs.

MSR, combined with the clustering system CLUTO (Karypis, 2002), groups the list of new lemmas into clusters of semantically related lemmas (50-200 in each). MSR and clustering introduce errors, and one lemma can represent several LUs, so flaws in the final clusters are inevitable. Nevertheless, each cluster represents about 2-3 different domains. Editors are next assigned clusters of lemmas to work on. This division of work, supported by *WordnetLoom*, enables them to concentrate on a limited number of semantic domains.

The extracted knowledge source are next delivered to the *WordnetWeaver*, a subsystem of *WordnetLoom*. For each new lemma, *WordnetWeaver* generates suggested LUs and presents them visually as subgraphs of the existing hypernymy structure. Editors are not limited by the suggestions: they can freely edit the wordnet.

There is even more support for editors, a recent addition to *WordnetWeaver*: automatically extracted examples of LU uses, produced by a system for unsupervised Word Sense Disambiguation (WSD) called *LexCSD* (Broda, 2011). *LexCSD* first identifies potential senses of a lemma by clustering its occurrences in the corpus. Next, for each cluster the most representative use is selected. For each new lemma, the generated examples are presented in the bottom part of the *WordnetWeaver* screen. While not all senses are automatically extracted, the size of the corpus (1.2 billion tokens) and the variety of texts and genres mean that the examples often include senses not covered by the existing dictionaries.

There were several stages of the expansion of *plWordNet* 1.0 toward 2.0. There were three stages devoted to nouns, considering that this is the category perhaps most important for potential applications of *plWordNet*. There were some 26000 new lemmas on the lists, but the final number of lemmas added was much higher. The editors included many synsets or even hypernymy subgraphs not on the extracted list; though frequency considerations dominate the expansion process, we decided that it is better not to leave gaps in the new portions, because they might later be overlooked. The third stage saw some 9000 verb lemmas extracted from the corpus, some 13500 eventually added to *plWordNet* (more than 26500 new

LUs). The *plWordNet* statistics after the first three stages appear in Table 1. A new *plWordNet* is published every three months on the Web (www.plwordnet.pwr.wroc.pl) along with detailed statistics.

5 Lessons Learned

Semi-automatic wordnet-creation methods are far from producing results which would be acceptable without almost any human control. Nevertheless, they proved very useful: by “digging” into a very large corpus, they greatly helped increase the efficiency of the process and the coverage. It must be noted that a strict corpus-based procedure would almost certainly lead to many omissions clear to the native speaker. That is why we ask the editors to add units which they find obviously missing: a linguist, supported by a dictionary, can rather easily spot such lacunae.

The construction of the verb hypernymy structure benefitted from our verb classification. The verb class and the aspect are not elements of the relation-based description, but we refer to them in the definitions of relations. Both have influenced the relation system and became indirectly part of the description (Maziarz et al., 2011b).

The *plWordNet* structure is crucially shaped by the constitutive relations. They include no derivationally motivated relations, but relations which originate from derivational associations help differentiate LUs much more accurately. They link LUs, not word forms, and quite often only two particular LUs derived from the same lemmas are linked. As an example, consider the word “kometka”. There are two LUs, *kometka* 1 ‘badminton’ and *kometka* 2 ‘small comet’, but the relation *markedness:diminutivity* can link only *kometka* 1 to the LU *kometka* 1 ‘comet’.

6 More to Come

The present version of *plWordNet* is already large, but several expansion stages are still required to achieve the shape planned for version 2.0. First, we will create semi-automatically the hypernymy structure of nouns derived from verbs.⁵ The structure will be based on the existing verb hypernymy structure. We want to add derivatives of the already described verb derivative bases. The analysis of a sample helped estimate that only 5%

⁵Polish verbal nouns are similar to English gerunds, but they function more as independent nominal LUs.

verbs will not have corresponding gerunds. Most of the verb hypernymy structure should be easily transferred to the noun component. The difficulty may be in merging the structures with the existing ones. Some gerunds were described in *plWordNet* 1.0. Also, verb hypernymy is more ‘bushy’, while gerundial structures will be mostly linked to the upper parts of the noun hypernymy structure. We expect to add some 20000 new noun LUs.

For the adjective component, a system of relations must be developed, perhaps inspired by a most interesting system in the Portuguese *WordNet.PT* (Marrafa and Mendes, 2006). We plan to add ≈ 15000 adjective LUs.

The Polish derivational mechanisms are relatively regular and very productive. We are working on automatic recognition of derivational relations with a tool trained on derivational pairs already described. The tool, applied to a long list of Polish lemmas, will identify derivatives and derivative bases not yet present in *plWordNet*. *WordnetWeaver* will be expanded to facilitate semi-automatic addition of LUs based the generated results. We expect to add 5000-8000 LUs.

The development of *plWordNet* has been monolingual all along, but mapping *plWordNet* synsets to *Princeton WordNet* synsets has always been an important element of our long-term plans. The process, slated for the year 2012, should provide mapping for at least 40000 of *plWordNet*'s noun LUs at the higher levels. We plan to perform the mapping in two iteratively repeated phases: first, verify and correct a selected part of the hypernymy structure (from the monolingual perspective), and then build a mapping for exactly the same hypernymy subgraph. We envisage some form of semi-automatic approach based on existing resources and methods. We expect that some new LUs can be added during verification and correction, so the final size of *plWordNet* at the end of the current project should reach 140000-150000 LUs in more than 100000 synsets.

Acknowledgments

Co-financed by the Polish Ministry of Education and Science, Project N N16 068637, and the European Innovative Economy Programme project POIG.01.01.02-14-013/09.

References

Mirosław Bariko, editor. 2000. *Inny słownik języka polskiego PWN*, volume 1-2. Wydawnictwo

	Nouns	Adjectives	Verbs	All
Lemmas	46746	3404	17009	67159
Lexical units	59467	4724	31133	95324
Synsets	44192	2791	21078	68061
Monosemous lemmas	37854	2582	9913	50349
Polysemous lemmas	8892	822	7096	16810

Table 1: *plWordNet 1.5* in numbers, August 2011.

- Naukowe PWN, Warszawa.
- Bartosz Broda. 2011. Evaluating lexicographer controlled semi-automatic word sense disambiguation method in a large scale experiment. *Control and Cybernetics*, ???:??–?? (to appear).
- D.A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge, UK.
- Magdalena Derwojedowa, Stanisław Szpakowicz, Magdalena Zawiśawska, and Maciej Piasecki. 2008. Lexical units as the centrepiece of a wordnet. In M. A. Kłopotek, A. Przeźiórkowski, S. T. Wierzchoń, and K. Trojanowski, editors, *Proc. 16th Int. Conf. on Intelligent Information Systems*, pages 351–358.
- Stanisław Dubisz, editor. 2004. *Uniwersalny słownik języka polskiego [a universal dictionary of Polish], electronic version 1.0*. PWN.
- Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. The MIT Press.
- George Karypis. 2002. CLUTO a clustering toolkit. Technical Report 02-017, Department of Computer Science, University of Minnesota.
- Claudia Kunze and Lothar Lemnitzer. 2002. GermaNet – representation, visualization, application. In *Proc. LREC 2002, main conference*, volume V, pages 1485–1491.
- C. Kunze. 1999. Semantics of verbs within GermaNet and EuroWordNet. In E. Kordoni, editor, *Workshop Proc. 11th European Summer School in Logic, Language and Information*, pages 189–200.
- Michał Marcińczuk and Maciej Piasecki. 2011. Statistical proper name recognition in polish economic texts. *Control and Cybernetics*, ???:?? (to appear).
- Palmira Marrafa and Sara Mendes. 2006. Modeling Adjectives in Computational Relational Lexica. In *Proc. COLING/ACL 2006 Main Conf. Poster Sessions*, pages 555–562, Sydney, Australia.
- Marek Maziarz, Maciej Piasecki, Joanna Rabiega-Wisniewska, and Stanisław Szpakowicz. 2011a. Semantic Relations among Nouns in Polish WordNet Grounded in Lexicographic and Semantic Tradition. *Cognitive Studies*, 11:161–181. www.eecs.uottawa.ca/~szpakk/pub/Maziarz_et_al_CS2011a.pdf.
- Marek Maziarz, Maciej Piasecki, Stanisław Szpakowicz, Joanna Rabiega-Wisniewska, and Bożena Hojka. 2011b. Semantic Relations Between Verbs in Polish Wordnet 2.0. *Cognitive Studies*, 11:183–200. www.eecs.uottawa.ca/~szpakk/pub/Maziarz_et_al_CS2011b.pdf.
- George A. Miller and Florentina Hristea. 2006. WordNet nouns: Classes and instances. *Computational Linguistics*, 32(1):1–3.
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-Line Lexical Database. *Int. J. of Lexicography*, 3(4):235–244.
- Ian Niles and Adam Pease. 2001. Towards a Standard Upper Ontology. In *Proc. Int. Conf. on Formal Ontology in Information Systems (FOIS-2001)*, pages 2–9, New York, NY. ACM.
- Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wrocław University of Technology Press. www.eecs.uottawa.ca/~szpakk/pub/A_Wordnet_from_the_Ground_Up.zip.
- Maciej Piasecki, Michał Marcińczuk, Adam Musiał, Radosław Ramocki, and Marek Maziarz. 2010a. WordnetLoom: a graph-based visual wordnet development framework. In *Proc. Int. Multiconf. on Computer Science and Information Technology - IMCSIT 2010, Wista, Poland, October 2010*, pages 469–476.
- Maciej Piasecki, Stan Szpakowicz, and Bartosz Broda. 2010b. Toward plWordNet 2.0. In P. Bhattacharyya, C. Fellbaum, and P. Vossen, editors, *Proc. 5th Global Wordnet Conf.*, pages 263–270. Narosa Publishing House.
- Maciej Piasecki, Roman Kurc, and Bartosz Broda. 2011. Heterogeneous Knowledge Sources in Graph-based Expansion of the Polish Wordnet. In *Proc. 2nd Asian Conf. on Intelligent Inform. and Database Systems*, LNAI 6591. Springer. (to appear).
- Piek Vossen. 2002. EuroWordNet General Document Version 3. Technical report, Univ. of Amsterdam.

NOUS IN ODIA: AN ONTOLOGICAL PERSPECTIVE¹

Panchanan Mohanty

Centre for Applied Linguistics and Translation Studies
University of Hyderabad
Hyderabad – 500046 (India)
E-mail: panchananmohanty@gmail.com

Abstract

Neither the lexicographers nor the grammarians of Odia have considered the feature ‘animacy’ important in their dictionaries and grammar books respectively. Most probably they have not found it useful for their work. That is why, they have never classified Odia nouns into categories like [+animate] and [-animate]. As a result, no Odia dictionary has marked the nouns for animacy nor has any Odia grammar discussed its role in the language. But this paper demonstrates that such a categorisation is necessary for both the lexicographers as well as the grammarians dealing with Odia, because the use of this ontological information will make the dictionaries more suitable for NLP purposes and also help to produce more fine-grained grammatical descriptions.
(Key words: ontology, animacy, case-suffix attachment, subject-verb agreement, Odia)

1 Introduction

Ontologies, in general, refer to knowledge bases and “... are commonly defined as specification of shared conceptualizations. Intuitively, the conceptualization is the relevant informal knowledge one can extract and generalize from experience, observation

or introspection.” Prévot et al. (2010:3) But it has been discussed in different ways and form different perspectives in various disciplines to achieve each one’s specific goals. Hence, many ontologies are possible in a language, and there is no consensus among scholars as to what exactly ontology is.

As a result, the ontological discussions in different disciplines have given rise to different traditions (Vossen 2003; Prévot et al. 2010). On the other hand, there is a broad agreement among scholars as to what ontology is meant for. According to Jansen (2008:173), “The task of ontology is to represent reality or, rather, to support the sciences in their representation of reality.”

Therefore, any discussion on ontologies has to focus on the categorization of our knowledge regarding the realities of the world. “Thus, it is more likely that knowledge of such naturally existing categories will put us in a position to constant systematic representations of that domain which have some degree of predictive power. If we can predict the way in which entities in a domain will behave under certain conditions, we are better able to understand that domain, interact with it,

and gain more knowledge about it.”
(Munn 2008:13)

With reference to NLP, Vossen (2003:473) argues that ontological information or knowledge is used for the following two purposes: It “... helps the system to make a structural analysis of language, e.g. to resolve PP attachments, correct spelling and syntax errors, improve speech recognition” and “... is used to do partial understanding, e.g. query search in information retrieval, document classification, automatic semantics, word-sense disambiguation.”

In this paper, I would like to focus on the first one and argue that animacy is an important ontological feature in Odia (earlier Oriya), a major Neo Indo-Aryan language and official language of the state of Odisha (earlier Orissa). The point to be noted is that animacy is quite important in language analysis “because essentially the same kinds of conceptual distinction are found to be of structural relevance across a wide range of languages.” (Comrie 1983:178). But there is hardly any serious work on animacy in the context of Indian languages.

Considering the role of animacy in Odia, its nouns can be grouped under two categories: [+animate] and [-animate]. This argument will be substantiated on the basis of attachment of certain case-suffixes which are sensitive to this ontological information. It may not be

out of place to mention here that Odia is different from its sister languages like Assamese and Bengali because the degree of animacy is lower in the latter than the former. The following example from Odia will drive home the point:

1. (a) jONE lokO jauci
“one man is going”
‘A man is going.’
(b) Onek lokO jaucanti
“many man are going”
‘Many men are going.’

But

2. (a) goTe kar jauci
“one car is going”
‘A car is going.’
(b) Onek kar jauci
“many car is going”
‘Many cars are going.’

Notice that 1(b) has a plural subject and a plural verb whereas 2(b) has a plural subject (the subject ‘car’ is modified by ‘many’) and a singular verb, because the former subject is [+animate] whereas the latter is [-animate]. This kind of difference in the verb is absent in Assamese and Bengali. Thus, categorization of nouns into [+animate] and [-animate] is an important characteristic in Odia, and that is why it is emphasized in this paper

2 Odia Case-suffixes

Let us now list the case-suffixes in Odia which are as follows:

Nominative	: Ø
Accusative	: ku, te, Ø
Instrumental	: dwara, dei, re
Comitative	: sOhO, sOhitO, sOn'ge, san'ge, san'gOre, sathire ² , bina, chORa
Dative	: paiM, lagi, joguM, sOkase
Ablative	: Tharu~Thu, ru, u

Genitive : kOrO, kO, rO, karO, ka

Locative : Thare ~ Thi, re, e

But out of all these only those which are sensitive to the animacy feature will be considered in this paper.

3 General Remarks

Odia does not have any nominative case suffix; therefore, it is marked /Ø/. Inanimate as well as indefinite nouns also do not normally take any accusative marker (hence, it is /Ø/) when they are used as grammatical objects. For example:

3. sita-Ø bOhi pORhila
“Sita book read”
‘Sita read a book.’
4. se bhatO-Ø khaila
“he rice ate”
‘He ate rice.’
5. tu baghO-Ø dekhilu
“you tiger saw”
‘You saw a tiger.’

Notice that the nominative subject /sita/ in (3), the inanimate object /bhatO/ in (4), and the indefinite object /baghO/ in (5) do not have any case-suffixes attached to them. Therefore, /Ø/ has been posited as the nominative case-suffix as well as an accusative case-marker in all these examples.

The accusative suffix /-te/ is used only with the first person singular pronominal oblique stem /mo/ ‘my’ and the second person non-honorific singular pronominal oblique stem /to/ ‘your’, whereas /-ku/ is used in all other cases.

For example:

mote / * mo-ku ‘to me’

to-te/ * to-ku ‘to you’

*ta-te / ta-ku ‘to him’

*amO-te / amO-ku ‘to us’

*semanOn'-te / semanOn'-ku ‘to them’

*ramO-te / ramO-ku ‘to Ram’

*sita-te / sita-ku ‘to Sita’

The genitive endings /-karO/ and /-ka/ can be added only to a few adverbs of time and place. Interestingly, these adverbs are also deictic in nature. The following examples will drive home the point:

eThi-karO	‘of here’
eThi-ka	‘of here’
aji-karO	‘of to-day’
aji-ka	‘of to-day’

On the other hand, /-kOrO/ and /-kO/ are always added to the honorific and/or plural oblique stems; but /-rO/ is added to those stems which are non-honorific and singular. For example³:

Plural	Honorific	Non-hon. & Singular
semanOn'-kOrO/	tan'-kOrO/	*ta-kOrO /ta-rO
*semanOn'-rO	*tan'-rO	‘his’ (non-hon.)
‘their’	‘his’ (hon.)	
bhaimanOn'-kO	bhain'-kO	*bhai-kO
/*bhaimanOn'-rO	/*bhain'-rO	/bhai-rO
‘of brothers’	‘of brother’s’	‘of brother’s’
	(hon.)	(non-hon.)

This distinction between /-kOrO/ and /-rO/ can be schematized in a tabular form as follows:

Genitive case-suffix	hon. and/or pl.
kOrO	+
kO	+
rO	-

4 Animacy in Odia

Though animacy plays a significant role in Odia, it has been completely neglected by the scholars of this language. That is

why, a very interesting facet of certain Odia case-endings has remained unnoticed till now. To be specific, the instrumental, ablative, genitive, and locative endings are clearly divisible into two categories: [+animate] and [-animate]. In other words, certain instrumental, ablative, genitive, and locative endings are added to [+animate] stems whereas the remaining instrumental, ablative, genitive, and locative endings are attached to [-animate] stems. Thus, they are mutually exclusive as will be made clear in what follows.

4.1 The Instrumental Suffixes

Out of the three instrumental endings, /-re/ is used with [-animate] stems, and the other two, such as /-dvara, -dei/ with [+animate] stems.

- 6. (a) muM cakOrO- $\left\{ \begin{array}{c} \text{dvara} \\ \text{dei} \end{array} \right\}$ ghOrO dhueili
“I servant- by house got washed”
‘I got the house washed by the servant.’
- (b) *muM cakOrO-re ghOrO dhueili
- 7. (a) muM churi-re seu kaTili
“I knife- with apple cut”
‘I cut an apple with the knife.’
- (b) *muM churi- $\left\{ \begin{array}{c} \text{dvara} \\ \text{dei} \end{array} \right\}$ seu kaTili⁴

Here 6(a) has a [+ animate] stem /cakOrO/ ‘servant’ and it is acceptable because /-dvara/ and /-dei/ are used with it. But attachment of /-re/ to the same stem in 6(b) has made it unacceptable. In 7(a) /-re/ has been added to [-animate]

/churi/ ‘knife’ and, therefore, it is acceptable. But when /-dvara, -dei/ are added to it in 7(b), the sentence becomes unacceptable. So we can argue that /-dvara, -dei/ are [+animate] and /-re/ is [-animate] case-suffixes.

4.2 The Ablative Suffixes

The ablative ending [-Thu] is an allomorph of /-Tharu/ and both are widely used in the spoken and written varieties of Odia respectively. Among all the ablative endings /-u, -ru/ are added exclusively to [-animate] stems. Again, /-u/ is attached mostly to stems that convey a deictic or directional meaning, e.g. /upOr-u/ ‘from above’, /tOL-u/ ‘from below’, /aR-u/ ‘from a direction’, /pOT-u/ ‘from a side’, /pakh-u/ ‘from a side’, etc. However, consider the following examples:

- 8. (a) muM ghOr-u asili
“I home-from came”
‘I came from home.’
- (b) *muM ta-u asili
“I him-from came”
‘I came from him.’
- 9. (a) muM kOTOkO-ru asili
“I Cuttack-from came”
‘I came from Cuttack.’
- (b) *muM ta-ru asili
“I him-from came”
‘I came from him.’

The cause of unacceptability of 8(b) and 9(b) is due to attachment of the [-animate] endings /-u, -ru/ to the [+animate] oblique stem /ta-/.

Then, /-Tharu~Thu/ can be added to both [+animate] as well as [-animate] stems. For example:

10. (a) se mo- $\left\{ \begin{array}{l} \text{Tharu} \\ \text{Thu} \end{array} \right\}$ bOhi nObO
 “he me- from book will take”
 ‘He will take a book from me.’
- (b) darjilin’ kOTOKO $\left\{ \begin{array}{l} \text{Tharu} \\ \text{Thu} \end{array} \right\}$ bOhut durO
 “Darjeeling Cuttack-from much distant”
 ‘Darjeeling is very far from Cuttack.’

Among all these endings, /-u/ is an organic one which Odia has inherited from its parent Prakrit. Between the other two, /-ru/ is the actual ablative ending whereas /-Tharu/ denotes comparison. For example:

11. (a) gOchO-ru nORia pORila
 “tree-from coconut fell”
 ‘A coconut fell from the tree.’
- (b) *gOchO- $\left\{ \begin{array}{l} \text{Tharu} \\ \text{Thu} \end{array} \right\}$ nORia pORila
12. (a) mo kOlOmO- $\left\{ \begin{array}{l} \text{Tharu} \\ \text{Thu} \end{array} \right\}$ ta kOlOmO bhOlO
 “my pen-from his pen good”
 ‘His pen is better than mine.’
- (b) *mo kOlOmO-ru ta kOlOmO bhOlO

11(b) is unacceptable due to the substitution of the real ablative ending /-ru/ by the comparison marker /-Tharu~Thu/ in it, and 12(b) is unacceptable as the sentence requires the comparison marker /-Tharu~Thu/, and not the ablative /-ru/. It should be noted here that /-Tharu~Thu/ are used not only as comparison markers, but also as ablative endings with the time and place adverbs.

For example:

13. (a) kOTOKO-ru darjilin’ bOhut durO
 “Cuttack-from Darjeeling much distant”
- (b) kOTOKO- $\left\{ \begin{array}{l} \text{Tharu} \\ \text{Thu} \end{array} \right\}$ darjilin’ bOhut durO
 ‘Darjeeling is very far from Cuttack.’

4.3 The Genitive Suffixes

Now let us consider the genitive endings /-kOrO/, /kO/, /rO/; because it has been pointed out above that the other two endings /-karO, ka/ are added exclusively to certain adverbs of time and place. But before that it should be mentioned that besides the genitive /-kO/ there is another /-kO/ which is a definitizer in Odia. For example:

14. (a) gOchOguRa dhire dhire mOrijauci
 “trees slowly slowly are dying”
 ‘The trees are dying out slowly.’
- (b) gOchOguRa-kO dhire dhire mOrijauci
 “the trees slowly slowly are dying”
 ‘The trees are dying out slowly.’

In 14(a) /gOchOguRa/ is [-definite] whereas in 14(b) with the attachment of /-kO/, /gOchOguRa-kO/ becomes [+definite]. If the genitive ending /-rO/ is added to this definitizer /-kO/, we get a compound morpheme /-kO-rO/ which is different from the genitive ending /-kOrO/. Compare the following sentences:

15. (a) ei pilaguRan’-kOrO bhagyO khOrap
 “these boys’ luck bad”
 ‘The luck of these boys is bad.’
- (b) ei pilaguRan’-kO bhagyO khOrap
 “these boys’ luck bad”
 ‘The luck of these boys is bad.’
16. (a) ei Tebulgura-kO-rO dam besi
 “these tables’ price much”
 ‘The price of these tables is high.’
- (b) *ei Tebulgura-kO dam besi

The difference between 15(a) and (b) is that the former has the genitive ending /-kOrO/ and the latter, another genitive ending /-kO/. Apparently, the same kind of difference seems to exist between 16(a) and (b). But it is significant that both 15(a) and (b) are perfectly acceptable, and though there is no problem with respect to 16(a), 16(b) is not acceptable. A careful examination will reveal that, unlike /-kOrO/ in 15(a), /-kO-rO/ of 16(a) consists of two morphemes, i.e. /-kO/ and /-rO/. Then, unlike /-kO/ of 15(b), /-kO/ of 16(b) is a definitizer; not a genitive ending. This has rendered 16(b) unacceptable.

From these examples and the discussion above, it is clear that out of the genitive endings /-kOrO, -kO, -rO/, the former two are added only to [+animate] stems that are [+plural] and/or [+honorific] and /-rO/ is added to either [-animate] stems or those which are both [+animate] and [+singular]⁵. Again, /-rO/ has to be obligatorily present if the stem is [-animate] whereas its presence is optional if the stem is [-animate]. The following examples are illustrative:

- 17. (a) eiTa amO-rO Tebul
“this our table”
‘This is our table.’
- (b) eiTa amO Tebul
“this our table”
‘This is our table.’
- 18. (a) eiTa amO Tebul-rO goRO
“This our table’s leg”
‘This is a leg of our table.’

- (b) *eiTa amO Tebul goRO⁶
“this our table leg”.

In 17(a), /-rO/ is overtly present whereas it is deleted in 17(b). On the other hand, because /-rO/ is used in 18(a) it is acceptable, and its deletion in 18(b) makes the sentence unacceptable. These observations provide further strength and support to our claim that if a noun stem is [-animate], /-rO/ has to be used obligatorily and its use is optional in the case of [+animate] stems.

4.4 The Locative Suffixes

Like the ablative /-u, -ru/, the locative /-e, -re/ get attached to the [-animate] stems only and never to [+animate] stems. For example:

- 19. (a) se ghOr-e Ochi
“he home-at is”
‘He is at home.’
- (b) *se ta-e Ochi
“he him-at is”
‘He is with him.’
- 20. (a) se kOTOkO-re Ochi
“he Cuttack-at is”
‘He is at Cuttack.’
- (b) *se ta-re Ochi
“he him-at is”
‘He is with him.’

Here /ghOrO/ ‘home, house’ in 19(a) and /kOTOkO/ ‘Cuttack (name of a city in Odisha)’ in 20(a) are [-animate]. That is why attachment of the locative suffixes /-e/ and /-re/ to them respectively does not create any problem. On the other hand, /ta/ ‘he (oblique)’ is [+animate] and, therefore, 19(b) and 20(b) have

become unacceptable due to the addition of /-e/ and /re/ respectively to it.

Then between /-Thare~Thi/, the latter is the abbreviated form of the former. Again, like /-Tharu~Thu/, /-Thare~Thi/ can be added to both [+animate] and [-animate] stems. The following examples are illustrative:

21. (a) mo- $\left\{ \begin{array}{l} \text{Thare} \\ \text{Thi} \end{array} \right\}$ TOn'ka pOisa kichi nahiM
 “me-at rupee paise some is not”
 ‘I have no money at all.’
- (b) ghOrO $\left\{ \begin{array}{l} \text{Thar}\bar{a} \\ \text{Thi} \end{array} \right\}$ TOn'ka pOisa kichi nahiM
 “home-at rupee paise some is not”
 ‘There is no money at home at all.’

5 Conclusion

From the above discussion it is clear that animacy plays a significant role in Odia and the case-suffixes in Odia can be grouped under two distinct categories, such as [+animate] and [-animate]:

Case suffixes	[animate]
Re	—
Dwara	+
Dei	+
U	—
Tharu~Thu	+,-
kOrO	+
kO [-pl., -hon.]	+
rO [-pl., -hon.]	—
E	—
re	—
Thare~Thi	+,-

Use of this categorization in the analysis of the Odia language will be helpful in writing better grammars and marking of this feature in the nouns will create dictionaries that will be better suited for

computational purposes. Again, incorporation of these findings in an NLP system will be very useful in an array of tasks, like machine translation, detection and automatic correction of morpho-syntactic errors, and automatic case-suffix attachment.

Notes

- I thank the two anonymous referees for their comments on an earlier version of this paper, but the usual disclaimers apply.
- In this paper [O] is used for the half-open rounded back vowel; [T, Th, R, Rh, N, L] for the voiceless unaspirated retroflex stop, voiceless aspirated retroflex stop, unaspirated retroflex flap, aspirated retroflex flap, retroflex nasal, and retroflex lateral respectively; [n'] for the velar nasal; and [M] for nasalization.
- The lexical meaning of /san'gO/ and /sathi/ both is ‘companion’ and these words possess the feature [+animate]. It is important to note that the animate nouns in Odia can take only the locative case-markers /- Thare~Thi/, and not /-re/. But /san'gO/ and /sathi/ are found here with the [-animate] suffix /-re/. So I want to claim that these two words do not carry the feature [+animate] here. They are devoid of their semantic content and have undergone the process of grammaticalisation. As a result, the [-animate] locative suffix /-re/ has been attached to them.

4. According to the homorganic nasal rule in Odia, n → n' /-k.
5. Though such a sentence may occasionally be found in stylized Odia, it is considered unacceptable in both the written and the spoken varieties.
5. The only exceptions to this description are /ame/ ‘we’ and /tOme~tume/ ‘you (pl.)’ which, though [+animate], take /-rO/ in the oblique case, e.g. /amO-rO/ and /tOmO-rO~tumO-rO/. These exceptions provide an interesting piece of evidence that /ame/ and /tOme~tume/, now the plural forms of /muM/ ‘I’ and /tu/ ‘you (sg.)’ respectively, were certainly honorific singular first person and honorific singular second person pronouns at an earlier stage of Odia. The Old Odia texts are full of such examples.
6. A sentence like 18(b) will, of course, be quite acceptable if /Tebul/ and /goRO/ are used together, to form a compound /TebulgoRO/.
- Munn, Katherine 2008. Introduction: What is ontology for? In: *Applied Ontology: An Introduction*, ed. by Katherine Munn and Barry Smith. Frankfurt: Ontos Verlag, pp.7-19.
- Prévet, Laurent; Chu-Ren Huang; Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, and Alessandro Oltramari 2010. Ontology and the lexicon: a multidisciplinary perspective. In: *Ontology and the Lexicon; A Natural Language Processing Perspective*, ed. by Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari, and Laurent Prévet. Cambridge: Cambridge University Press, pp. 3-24.
- Vossen, Piek 2003. Ontologies. In: *The Oxford Handbook of Computational Linguistics*, ed. by Ruslan Mitkov. Oxford: Oxford University Press, pp. 464- 482.

References

- Comrie, Bernard 1983. *Language Universals and Linguistic Typology: Syntax and Morphology*. Oxford: Basil Blackwell. (first published in 1981)
- Jansen, Ludger 2008. Categories: The top-level ontology. In: *Applied Ontology: An Introduction*, ed. by Katherine Munn and Barry Smith. Frankfurt: Ontos Verlag, pp. 173-196.

Ontology of Sanskrit Wordnet: Nouns and Verbs

Sanghamitra Mohanty

North Orissa University

Baripada, Orissa, India – 751004

sangham1@rediffmail.com

K.P.Das Adhikary

S.B.S.S. Mahavidyalaya

Paschim Midinapur, India – 721128

krishnapada06@rediffmail.com

Abstract

Sanskrit is the mother Language of all the Indian Languages. This is the Oriental Language of India. It is unique for its strong base of grammar and rigidity. Many ancient and valuable documents *Veda* (वेद), *VedAnta* (वेदान्त) and *Upanishads* (उपनिषद्) are written in this language. To have an access to such manuscripts knowledge of Sanskrit language is essential. Sanskrit Wordnet is an effort to help people to learn the language and make a study of the knowledge available through it. This wordnet is developed based on *Navya Nyaya* (नव्यन्याय) Philosophy and Paninian Grammar. During the development of this wordnet the architecture of this wordnet contains features like Etymology and Analogy besides the normal features like Synonym, Hypernym and Antonym. *Navya Nyaya* Philosophy provides the scope for such extra analysis as it deals with the construction of a word in Sanskrit Language.

1 Introduction

Indian Languages are derived from the Oriental Sanskrit language and being influenced by *Prakrit* and *Pali* the other two ancient languages of India. This oriental language has a systematic and technically strong base of grammar and thus plays a vital role on the modern Indian languages. Understanding this language needs a strong base of grammar and concept of the objects. To cater the needs of this language we are developing a Sanskrit Wordnet, which can help one to have an access to all other languages for different types of Natural Language Processing (NLP).

A word has a great importance in the field of NLP. Information like meaning or Part-of-Speech can be obtained from a Lexical resource like Dictionary while a Wordnet which is also a special type of Lexical Resource can provide many more information like Synonymy, Atonymy, Hypernymy, Etymology and Analogy

besides some definition and examples of its use in the language. This provides the classification and the origin of its formation too.

To design a robust wordnet the Ontological analysis of the words has been opted at the initial stage as it provides path to proceed for the use of the words in many form of NLP. The *Navya-Nyaya ShAstra* (नव्यन्यायशास्त्र) of Bhattacharya Jayanta (Bhattacharya Jayanta Bhattacharya, 1969), “Nyayamanjari” (न्यायमञ्जरी) and Paninian Grammar (Bhattacharya V. N. 1884) help in analysing the base structure or (*PrakritirUpa*) (प्रकृतिरूप) of any word in Sanskrit language. This helps to get the Universal Concept level knowledge of any word. Some words which are Indian culture specific give a clear idea of the concept which explains and disambiguates the meaning at the universal level. As Sanskrit language is having universal importance with respect to the conceptual knowledge exchange through Natural Language Processing techniques, the development and use of Sanskrit Wordnet in the field of Universal Networking of Languages has a greater role to play.

2 Ontology of Sanskrit Wordnet

Ontology is expressed through entities, ideas, and events, along with their properties and relations in both Computer Science and in Philosophy for knowledge representation of knowledge in conceptual form. In a Wordnet when a word is encountered it is expected that besides the meaning and Part-of-Speech the Word may also provide other Extended Lexical information like Synonymy, Hypernymy, Atonymy, Etymology and Analogy of the word. For all these Ontological analysis of the word is quite essential. In such type of analysis the representation as per *Navya-Nyaya ShAstra* (नव्यन्यायशास्त्र) and Paninian Grammar are with technical justification and is more informative to get the base-structure (*PrakritirUpa*) (प्रकृतिरूप)

Bhattacharya Gadadhara (Bhattacharya et.al 1995), “Vyutpattivada” (व्युत्पत्तिवाद) of all Sanskrit words than that of a normal dictionary relevant to the NLP.

Defining the term Ontology it is the specific conceptualization of any word. Since it provides the concept as per philosophy or the logical representation so can be used or interpreted universally and helps in universal knowledge sharing. The formal definition of ontological commitment can be mentioned as a set of definitions of formal vocabulary with properties of knowledge sharing. Dealing with semantic web analysis an ontological commitment is an agreement to use a vocabulary like asking a query or making some assertive statements to get consistent meaning as per the theory of the ontology for knowledge exchange. It also helps for the Pragmatic analysis with respect to NLP. Ontologies are to describe ontological commitments for a set of agents so that they can communicate about a domain without operating on a globally shared principle.

The ontology of the words with respect to their base structure as per the lexical recognition are described below through Figure 2.1 to 2.5 which provides the idea to the issue of designing this computational lexicon the Wordnet for Sanskrit. This is done as per the theory of the famous *Navya-Nyaya* (नव्यन्येय) specialist *Navya-NaiyAayika* (नव्यनैयायिक), Visvanathanyaya Panchanana Bhattacharya (विश्वनाथन्याय-पञ्चनन-भट्टाचार्य) (16th century A.D.). He mentioned 8 methods for confirming syntactic and semantic information of any Sanskrit word which is coded in the following verse.

शक्तिग्रहं व्याकरणोपमान- कोशापत्वाक्यात् व्यवहारतश्च।
वाक्यस्य शेषाद् विवृतेर्वदन्ति सन्धिध्यतः सिद्धपदस्य तृद्वा: ॥

*shaktigraham vyAkaraNopamAna
koshAptavAkyAt yyavaHAratashcha /
vAkyasya sheshhAd vivRtervadanti
sannidhyatah siddhapadasya vRddhAh //*

Those eight methods are i) grammar, ii) analogy, iii) dictionary iv) statement of knowledgeable persons (logic), v) usage (example), vi) supplementary statement (Explanation), vii) paraphrase and viii) contiguity of well known words. A Word when selected for incorporation to the Wordnet of Sanskrit is first analysed for its grammar and meaning and

then the other ontological analysis are done for adding the logic and other features with respect to NLP.

2.1 Grammar

The first important item Grammatical Information consists of mainly four parts: Nominal (*nAmagata*) (नामगत) and Verbal (*kriyAgata*) (क्रियगत), which has further, sub-items. In this lexical analysis if a word comes under the nominal form it must either be Noun (*visheshhya*) (विशेष्य) or Adjective (*visheshhaNa*) (विशेषण) or Pronoun (*sarvanAma*) (सर्वनाम) or Particle (*avyaya*) (अव्यय).

Again, if a word satisfies the second level of grammatical information, then it must come under seven inflections (*sapta-vibhakti*) (सप्तविभक्ति). According to Paninian Grammar, *sup-vibhakti-s* (सुप्विभक्ति) are being known as *subanta* (सुबन्त). By nature they are known as suffix and being used after the nominal. Again, each inflection contains three numbers namely, singular, dual and plural. (Verbal words also contain such kinds of number). By this division *sup-vibhaktis* (सुप्विभक्ति) are of 21 types. Panini technically represents 21 varieties of *sup-vibhakti* (सुप्विभक्ति) as described in the following Table No.1. It is also to be noted that though *avyaya* (अव्यय) (particle) does not possess any inflectional suffix, yet it implies the meaning of suffix.

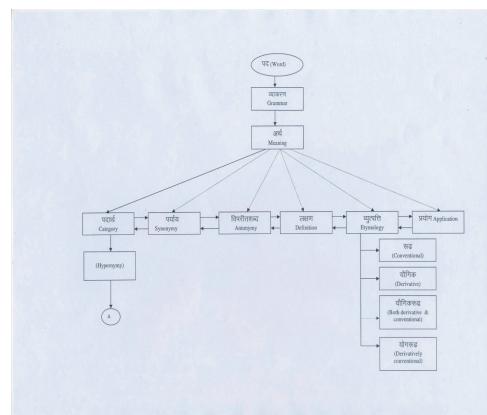


Figure 2.1 Analysis of a word to find its category

A word when selected can be analysed through the above processes to find its category through grammatical properties. If it comes broadly as noun or verb or pronoun then the appropriate process are to be conducted for the analysis. If it is a noun or a verb Figure 2.2 nar-

rates the steps to be followed to get the exact category.

(yatra ca kartR-bhinna-nishTham tatra parasmai-padam) [Panini Sutra⁴: 1-3-72/ 1-4-100/3-

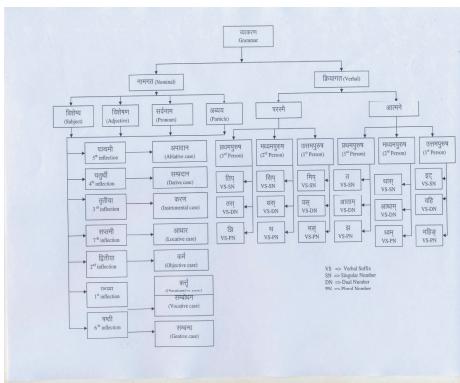


Figure 2.2 Analysis of a word for its category (Nominal or Verbal)

	Number		
	Singular	Dual	Plural
Inflections	1 st <i>su</i> (सु)	<i>au</i> (आौ)	<i>jas</i> (जस)
	2 nd <i>am</i> (अम्)	<i>aut</i> (आौट्)	<i>shas</i> (शस्)
	3 rd <i>TA</i> (टा)	<i>bhyAm</i> (भ्याम्)	<i>bhis</i> (भिस्)
	4 th <i>Nge</i> (ঁে)	<i>bhyAm</i> (ভ্যাম্)	<i>bhyas</i> (ভ্যস্)
	5 th <i>Ngasi</i> (ঁসি)	<i>bhyAm</i> (ভ্যাম্)	<i>bhyas</i> (ভ্যস্)
	6 th <i>Ngas</i> (ঁস্স)	<i>os</i> (আোস)	<i>Am</i> (আম)
	7 th <i>Ngi</i> (ঁও)	<i>os</i> (আোস)	<i>sup</i> (সুপু)

Table-1: 21 Nominal Inflections (सुबन्तप्रत्यय) for Nominal Words

In the second part of the diagram (Figure 2.2) the use of parasmai (परस्मै) and Atmane (आत्मने) for determining the nature of verbal words. By grammatical nature any verbal word can be either of parasmai (परस्मै) or Atmane (आत्मने). A root word (क्रियापद) whose verbal-result (*kriyAphala*) belongs to the agent (*kartR*) (कर्तु) only, then the word is called Atmane (आत्मने) (क्रियाफलं कर्तुनिष्ठं तत्र आत्मनेपदम्) (*kriyAphalam kartR-nishTha tatra Atmane-padam*). Unlikely, in case of parasmai (परस्मै) the result of a root word does not belong to the agent (यत्र च कर्तुनिष्ठ-निष्ठं तत्र परस्मै)

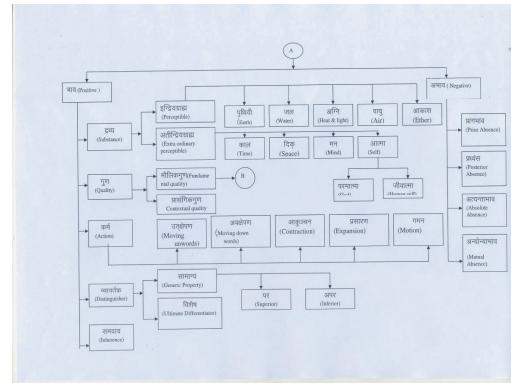


Figure 2.3 Ontological Analysis for any Word (Part – A)

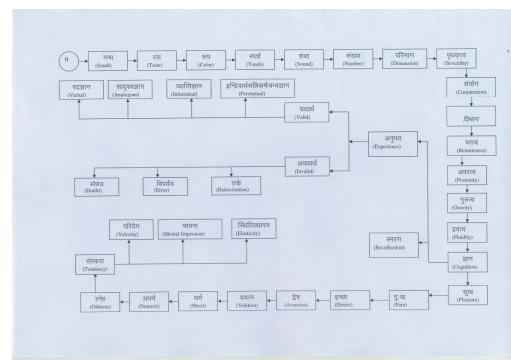


Figure 2.4 Ontological Analysis for a Word (Part – B, continuation of Part – A)

2-106/1-3-12). For the identification of parasmai and Atmane words Panini technically develops 18 suffixes (तिङ्गन्त) namely, *tip*, *tas*, *jhi* etc. These suffixes also contain further classification like number and person. They have been mentioned in the following Table No.2 and Table No. 3.

Number	3 rd Person	2 nd Person	1 st Person
Singular	<i>Tip</i> (তিঙ্গ)	<i>Sip</i> (সিঙ্গ)	<i>Mip</i> (মিঙ্গ)
Dual	<i>Tas</i> (তস)	<i>Thas</i> (থস)	<i>vas</i> (বস)
Plural	<i>Jhi</i> (ঁঁঁ)	<i>Tha</i> (ঁঁ)	<i>Mas</i> (মস)

Table 2. Parasmai is having the above 9 Suffixes.

Number	3 rd Person	2 nd Person	1 st Person
Singular	Tip (तिप्)	Sip (सिप्)	Mip (मिप्)
Dual	Tas (तस्)	Thas (थस्)	vas (वस्)
Plural	Jhi (झि)	Tha (ঝ)	Mas (মস্)

Table 3. *Atmane* is having the above 9 Suffixes.

3 Design of the Sanskrit Wordnet

The architecture of the database design used for English Word-Net by Fellbaum et.al. (Fellbaum 1999), Hindi Word-Net by Jha et.al. (Jha 2001), and Oriya Word-Net by Mohanty et.al. (Mohanty et. al. 2002), Sanskrit Word-Net by Mohanty et.al (Mohanty et.al. 2002) and Kulkarni et.al. (Kulkarni et.al. 2009) are also followed for this Word-Net. Along with it, we have used the *Navya-NyAya* (नव्यन्याय) Philosophy for a detailed analysis of a word. In this Word-Net we have analysed the category of the word first. In the next phase the synonymy, antonymy, hypernymy etc as per the Figures 2.1-2.4 are incorporated. The specialty of this wordnet is the incorporation of analogy and etymology as features.

3.1 Etymology.

Etymology has also important role with respect to the present Sanskrit Wordnet. The main purpose in organizing the items in the *Sanskrit Wordnet* is to determine the original construction of the word, which clearly describes the construction of the primary word (*mUlashabda*) (मूलशब्द) by base (*prakRti*) (प्रकृति) or *prAtipadika* (प्रातिपदिक) and suffix (*pratyaya*) (प्रत्यय). The *pratyayas* cited here don't include verbal inflection (विभक्ति) and nominal inflection (सुव्वन्त). Etymology is of four kinds namely, i) derivative (*yaugika*) (यौगिक), ii) conventional (*rUDha*) (रूढ), iii) derivatively conventional (*yogarUDha*) (योगरूढ) and iv) both derivative and conventional (*yaugika-rUDha*) (यौगिकरूढ). The compound words (*samAsa-sabda*) (समासशब्द) are although a construction of more than one primary word, yet it can be included in either of the *yaugika* (यौगिक) or *yogarUDha* (योगरूढ) or in *yaugika rUDha* (यौगिकरूढ) which are under further modification.

The architecture of the database design used for English Word-Net by Fellbaum et.al.⁴, Hindi Word-Net by Jha et.al.,⁵ and Oriya Word-Net by Mohanty et.al.⁶ are also followed for this Word-Net. Along with it, we have used the *Navya-NyAya* (नव्यन्याय) Philosophy for a detailed analysis of a word. In this Word-Net we have analysed the category of the word first. In the next phase the synonymy, antonymy, hypernymy etc are incorporated. The specialty of this wordnet is the incorporation of analogy and etymology as features.

3.1 Etymology.

Etymology has also important role with respect to the present Sanskrit Wordnet. The main purpose in organizing the items in the *Sanskrit Wordnet* is to determine the original construction of the word, which clearly describes the construction of the primary word (*mUlashabda*) (मूलशब्द) by base (*prakRti*) (प्रकृति) or *prAtipadika* (प्रातिपदिक) and suffix (*pratyaya*) (प्रत्यय). The *pratyayas* cited here don't include verbal inflection (विभक्ति) and nominal inflection (सुव्वन्त). Etymology is of four kinds namely, i) derivative (*yaugika*) (यौगिक), ii) conventional (*rUDha*) (रूढ), iii) derivatively conventional (*yogarUDha*) (योगरूढ) and iv) both derivative and conventional (*yaugika-rUDha*) (यौगिकरूढ). The compound words (*samAsa-sabda*) (समासशब्द) are although a construction of more than one primary word, yet it can be included in either of the *yaugika* (यौगिक) or *yogarUDha* (योगरूढ) or in *yaugika rUDha* (यौगिकरूढ) which are under further modification.

4 Examples from Sanskrit Wordnet

NOUN in Sanskrit wordnet:

The Noun समुद्रम् (Sea) has 15 senses in Sanskrit-Wordnet---

Sense 1:- समुद्रम् (Sea) भूपृष्ठस्थ-सर्ववृहत्कारजलसमूहः

⇒ आद्लाइटक्महासागरः युरोपात् अमेरिकां विभजते ।
बंगोपसागरः भारतमहासागरं संयुज्यते । सागरः भूभागं प्रायः

1. Synonyms of समुद्र (Sea)

उऊ आद्लाइटक्महासागरः युरोपात् अमेरिकां विभजते । बंगोपसागरः भारतमहासागरं संयुज्यते । सागरः भूभागं प्लावयति । नौकया जनाः सागरं तरन्ति । मेरुसागरः सूर्यकिरणेन द्रवति । अकारान्तः चैतेष्विशेषाः प्रथमः जातिवाचकविशेषाः तिप्, तस्, झि, ता, आताम्, झ



Figure 4.1 Snapshot of the Noun समुद्रः (Sea)

- Sense 1:- समुद्रम् (Sea) भूपृष्ठस्थ-सर्ववृहत्क्षारजलसमूहः
Sense 2 :- अब्धि, Sense 3 :- अङ्गुष्ठः, Sense 4 :- पारावार,
Sense 5 :- सरित्पति, Sense 6 :- उदन्चान्,
Sense 7 :- उदधि, Sense 8 :- सिंधुः, Sense 9 :- सरस्वान्,
Sense 10 :- सागर, Sense 11 :- अर्णव, Sense 12 :- रन्नाकर, Sense 13 :- जलनिधि, Sense 14 :- यादपति,
Sense 15 :- अपारपाति
2. Antonymy: Does not exist.
3. Hypernymy: नौमितिक
4. Category: द्रव्यम्
5. Etymology of समुद्रम् (sea) is योगिकरूढः(सम्-उद्-रा-क)
6. Analogy of समुद्रम् (Sea) does not exit in SanskritNet---
Sense 1 :- सरित्, Sense 2 :- हिम, Sense 3 :- करक
7. Definition: भूपृष्ठस्थ-सर्ववृहत्क्षारजलसमूहः
8. Application:
Sense 1 :- विभाजन, Sense 2 :- संयोजन, Sense 3 :- प्लावन, Sense 4 :- तारण, Sense 5 :- द्रवण
9. Example:
Sense 1 :- आदूलाण्टिकमहासागरः युरोपात् अमेरिकां विभजते। बंगोपसागरः भारतमहासागरं संयुजते। सागरः भूमारं प्लावयति। नीकया जातः सागरं तरचति। मेरुसागरः सूर्यीकरणेन द्रवयति। अकारान्तः पूँलिङ्गः प्रथमः जातिवाचकविशेष्यः तिपि
Sense 2 :- तस, Sense 3 :- द्वि, Sense 4 :- ता, Sense 5 :- आताम्, Sense 6 :- झ
10. Morphology: null

VERB in Sanskrit wordnet:

The Verb खाद् (Eat) has 6 senses in Sanskritnet-
1. Synonyms:

Sense 1:- खाद् (Eat)

उक्त शिशुः गृहे चमसेन अन्नं खादति। काकः वृक्षे चञ्च्या कीटं खादति। धेनुः प्रान्तरे मुखेन तृणं खादति।

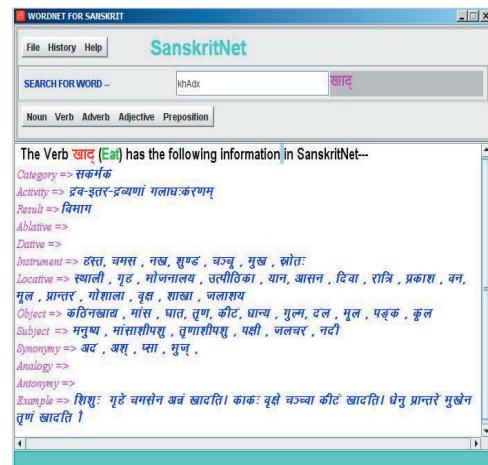


Figure 4.2 Snapshot of the verb खाद् (Eat)

- Sense 2 :- अद्, Sense 3 :- अश्, Sense 4 :- या, Sense 5 :- भुज्
2. Antonymy खाद् (Eat): Does not exist.
3. Result for the खाद् (Eat) is: विभाज
4. Instrument:
Instrument form of the verb खाद् (Eat) is—
Type 1 => हस्त, Type 2 => चमस, Type 3 => नख,
Type 4 => शुण्ड, Type 5 => चञ्चू,
Type 6 => मुख, Type 7 => खोतः
5. Locative:
Locative form of खाद् (Eat) is --
Type 1 => स्थाली, Type 2 => गृह, Type 3 => भोजनालय, Type 4 => उत्पातिका, Type 5 => यात्रा
Type 6 => आसन ठनखादा, Type 2 => मा , Type 7 => विचार, Type 8 => रात्रि, Type 9 => प्रकाश,
Type 10 => वन, Type 11 => मूल, Type 12 => प्रान्तर, Type 13 => गोशाला, Type 14 => वृक्ष,
Type 15 => शाखा, Type 16 => जलाशय,
6. Activity:
Activity of खाद् (Eat) is: द्रव-इतर-द्रव्यणां गलाधःकरणम्
7. Object:
The Object of खाद् (Eat) is---
Type 1 => कठिनखाद, Type 2 => मांस, Type 3 => घात, Type 4 => तृण, Type 5 => कीट,
Type 6 => धान्य, Type 7 => गुल्म, Type 8 => दल, Type 9 => मूल, Type 10 => पद्मक,
Type 11 => कूल,
6. Subject:
The Subject of खाद् (Eat) is---
Type 1 => मनुष्य, Type 2 => मांसाशीपशु, Type 3 => तुणाशीपशु, Type 4 => पश्ची,
Type 5 => जलचर, Type 6 => नदी,
7. Example:
The Example of खाद् (Eat) is--

Type 1 => शिशुः गृहे चमसेन अन्नं खादति ।

Type 2 => काकः वृक्षे चञ्च्चा कीटं खादति ।

Type 3 => धेनुं प्रान्तरे मुखेन तृणं खादति ।

9. Morphology:

The Morphology of खाद् (Eat) ----

Class (gaNa) => भवादि

Type (pada) => परस्मै

PRONOUN (*sarvanAma*) (सर्वनाम) (*avaya*) (अस्मत्) or Particle (*avyaya*) (अव्यय) in Sanskrit wordnet:

1. Grammatical information of *asmad*:

i. Nominal:

ii. *visheshhya*(विशेष्य), *sarvanAma*(सर्वनाम)

iii. Possesses seven inflections (सुप्-विभक्ति)

iv. Implies respective *karakas* (कारक)

2. Category:

i. *dravya* (द्रव्य) =>*AtmA*(आत्मा) =>

(*manusyAtmA*) (मनुष्यात्मा) => speaker

3. Antonym: i. (*hearer* in semantic level)

4. Analogy: i. *sva*(स्व) (self)

5. Etymology: *rUDha* (रुढ़)

svatantra-uccArayitA(स्वतन्त्र-उच्चारयिता)



Figure 4.3 Snapshot of the Pronoun (*sarvanAma*) (सर्वनाम) *avaya* (अस्मत्).

5 Conclusion

Sanskrit Wordnet developed by us is a specially designed lexical resource which can be of used in Natural Language processing extensively

as it provides comprehensive and systematic information about a word. The semantic of a word or a sentence offers same effect in all languages. We have made a detailed analysis of the *Navya Nyaya* philosophy and its use in the development of the Sanskrit Wordnet and have found that this philosophy provides a strong background for the wordnet. This wordnet can provide a background as an interlingua for any type of Cross Lingual Natural Language Processing as most of the Indian and Indo Aryan languages are developed from this Oriental Language. Since this wordnet is developed on the basis of fundamental concept which is true in situation it can be used to make a link with any of the Indo Aryan Languages too. Snapshot of different words gives a flavor of the Sanskrit Wordnet developed by us.

References:

Bhatta Jayanta, “Nyayamanjari” (न्यायमञ्जरी), Ed. K. S. Baradacarya, published by The Oriental Research Institute, Mysore, 1969, Vol II.

Bhattacharya Vishvanatha Nyayapanchanana , “Bhasapariccheda” (भाषापरिच्छेद), Published by Sr.Satinatha Bhattacharya ,9 Rajakrishna lane, Calcutta-6,1884.

Bhattacharya Gadadhara, “Vyutpattivada” (व्युत्पत्तिवाद), Ed. Venimadhav Sukla, Chauk-hambha, Sanskrit Sansthan. Varanasi – 221001, 1995.

Fellbaum C., “Word Net: An Electronic Lexical database”, The MIT Press, Cambridge, Massachusetts, London, England, 1999.

Jha S.K. et. al., “Hindi Wordnet”, Proceedings of the Workshop on Lexical Resources for Natural Language Processing, 2001, IIIT, Hyderabad.

Mohanty et. al., “Oriya Wordnet”, Proceedings of the 1st Global Wordnet Conference, 2002, Central Institute of Indian Language , Mysore.

Mohanty S., Santi P. K., Das Adhikary K. P. and Nayak S. N. “Making of a Sanskrit Wordnet”, ICUKL, 2002.

Kulkarni M. Bhattacharya P. “Verbal Roots in Sanskrit Wordnet”, Springer-Verlag, 2009.

Using WordNet to predict numeral classifiers in Chinese and Japanese

Hazel Mok Shu Wen, Gao Huini Eshley and Francis Bond

Linguistics and Multilingual Studies

Nanyang Technological University

haze0004@e.ntu.edu.sg, gaoh004@e.ntu.edu.sg, bond@ieee.org

Abstract

Most nouns must be modified by a numeral-classifier combination when quantified in classifier languages like Chinese and Japanese. In this paper, we present a method to generate numeral classifiers using Chinese and Japanese WordNets. We assign synsets from WordNet to each classifier by hand and use a modified algorithm to generate sortal classifiers based on semantic hierarchies. We obtained a generation score of 78.80% for Chinese and 89.84% for Japanese.

1 Introduction

Classifiers have long been an interest for many linguists. In many Asian languages like Japanese and Chinese, nouns often require numeral classifiers when they are quantified (Bond & Paik, 2000; Downing, 1996). This contrasts with English where count nouns can be modified by a numeral. In English, it is acceptable to say “two books”. In classifier languages, it is obligatory to use a numeral-classifier pairing, as in (1) and (2).

(1) Japanese: 2 冊 の 本

2-satsu-no hon
2-CL-ADN book
“2 books”

(2) Chinese: 两 本 书

liǎng běn shù
2 CL book
“2 books”

Numeral classifier languages typically lack plural markings. Greenberg (1972) has suggested that a classifier in these languages functions like a plural suffix in languages that require plural markings (cited in Downing, 1996). Hundius and Kölver (1983) argue that a classifier establishes immediate reference to individual objects.

The type of classifier used depends on the semantic features of the noun referent (Zhang, 2007). Some of these semantic categories in-

clude animacy and shape. For example, 只 *zhī* is commonly used as the classifier for counting “animals” in Chinese. However, there are also instances when nouns with similar properties use different classifiers (Guo and Zhong, 2005). For instance, 条 *tiáo*, which is used for long and thin objects like “ropes”, is also used as the classifier for “snakes”.

The Japanese distribution of classifiers is dependent on their referent classes: animates, concrete inanimates and abstract inanimates (Downing, 1996).

Bond and Paik (2000) identified five major types of classifiers that have different properties depending on the context. These are sortal, event, mensural, group and taxonomic.

Sortal classifiers are defined as those which classify the kind of noun they count (e.g. 辆 *liàng*, the classifier for “vehicle”). Mensural classifiers measure some property of the object denoted by the noun they modify (e.g. 米 *mǐ* “metre”). Event classifiers are used to count events (e.g. 次 *cì* “time”). Group classifiers refer to a set of individuals belonging to the type denoted by the noun (e.g. 组 *zǔ* “group”). Taxonomic classifiers exert a generic interpretation of the noun phrase that they modify (e.g. 种 *zhǒng* “kind”).

Understanding the classifier systems of these languages can help us to appreciate how the languages cover the hierarchy of meaning with the use of classifiers. In addition, it will provide us with some insight on how speakers of these languages view the world.

Classifier systems are usually very complex. They are also one of the more difficult aspects of grammar to acquire. Even native speakers may have difficulty using some of them. Furthermore, classifiers are often poorly translated. The wrong classifier may be used or left out altogether. Hence, we hope that this study can help to im-

prove the accuracy and efficiency of machine translation. The results of this study can also benefit learners of Japanese and Chinese by helping them retrieve the appropriate classifier when forming noun phrases.

This paper is structured as follows. Section 2 gives a brief overview of work that has been conducted in this area thus far. It also introduces the resources that we will be using for this study. Section 3 documents the methodology. Then, we present the results in Section 4 and discuss the significance of the results and the limitations faced in Section 5.

2 Background

There has been much more work done on analyzing classifiers than in generating them in natural language processing. One important study investigating the generation of classifiers in Thai was carried out by Sornlertlamvanich, Pantachat & Meknavin (1994). The authors proposed an algorithm for matching an appropriate classifier with a noun. Their study involved obtaining noun-classifier pairs from a tagged, word-segmented corpus. From the pattern of noun-classifier collocations, they determined the best representative classifier for each noun and semantic class. However, they did not include a detailed evaluation of the accuracy of their algorithm.

Bond and Paik (2000, 2001) presented a modified algorithm based on Sornlertlamvanich et al's (1994) work. This modified algorithm was used for associating classifiers with semantic classes in Japanese and Korean. It is able to handle nouns which belong to more than one semantic class. It does this by organizing the semantic classes according to the noun referent's most frequent use. The general idea is to assign the default classifier of the most typical semantic class to the noun.

Resources

There are 145,000 synsets for different parts of speech (nouns, adjectives, verbs, adverbs) in the Princeton WordNet of English v3.0 (PWN: Fellbaum 1998). The structure of WordNet allows one to see the relationship between words such as hypernyms (superordinates) and hyponyms (subordinates). It is often used for work in natural language processing.

The Japanese Wordnet (JWN: Isahara *et al* 2008), contains about 57,238 synsets based on the same lexical arrangement as PWN. This means that lexical units in the Japanese Wordnet were arranged according to their hierarchical

connections among words as well. However, the Japanese and English wordnets are not a direct copy of each other; for instance, there are Japanese synsets that are not found in the English wordnet and vice versa due to the uniqueness of both languages (Isahara *et al.*, 2009). One example is the concept of “rice”. Japanese makes a distinction between 米 *kome* “rice” and 御飯 *gohan* “cooked rice”. This distinction is not made in English, and therefore the English wordnet does not include a separate entry for the two senses.

The Wordnet used for Chinese is a bilingual Chinese-English Wordnet (CWN: Xu, Gao, Pan, Qu and Huang, 2008). It is a bilingual lexical database, which also uses the semantic hierarchy from WN. This Chinese-English Wordnet has more than 150,000 Chinese words. Each Chinese synset is linked to an English synset, which allows for useful cross-language information retrieval.

We used a 38,000 sentence Japanese-English-Chinese corpus, the NICT Multilingual corpus, (Zhang, Uchimoto, Ma and Isahara, 2005) based on the Kyoto text corpus. The corpus was created using Japanese sentences from Mainichi Newspaper and manually translated into Chinese and English.

3 Methodology

This section documents the steps taken in the study.

3.1 Categorisation of classifiers

We extracted 228 Japanese classifiers and 264 Mandarin Chinese numeral classifiers from the corpus. This was done by extracting anything tagged as classifier for part-of-speech (POS). For Japanese, we pulled out every word that was tagged with *meishi-setsubi-jousushi* “noun-suffix-classifier”. For Chinese, q.* was the POS for classifier.

These were sorted into the following categories: sortal, mensural, date and time, currency and not classifier. Sortal and mensural classifiers were defined as mentioned before. Date and time classifiers measure the span of days and time periods (such as 年 *nián* “year” and 秒 *miǎo* “second”). Currency classifiers are used to refer to a country’s currency (such as 美元 *měiyuán* “American dollar”). Lastly, nouns which had been paired with a numeral, but were in fact not

classifiers, were removed (such as 三页 *sānyè* “three pages” and 两餐 *liǎngcān* “two meals”).

3.2 Hand annotation of corpus

Distribution pattern	Example
classifier-no-noun	2匹の犬
(NUM)+CL no (NOUN)+	2-hiki-no-inu “2 of the dogs”
noun-no-classifier	犬の2匹
(NOUN)+ no (NUM)+CL	inu-no-2-hiki “2 dogs”
noun-ga/wo/mo/wa-classifier	犬が/を/も/は/2匹
(NOUN)+ ga/wo/mo/wa (NUM)+CL	inu-ga/wo/mo/wa-2-hiki “dogs, 2”

Table 1. Distribution pattern for Japanese

Distribution pattern:
(DET)? (NUM)+ <u>CL</u> ¹ (NOUN)+ ²

Table 2. Distribution pattern for Chinese

55 sortal classifiers were identified for Japanese and we extracted sentences containing noun phrases that are modified by those sortal classifiers. We did the same for Chinese. Chinese had more classifiers with 136 sortal classifiers identified in the previous step. Classifiers that appeared more than 100 times had their counts reduced. We used distribution patterns of classifiers and nouns to retrieve sentences with the numeral-classifier combination and the noun phrases. These distribution patterns were identified using language dependent patterns. Tables 1 and 2 show the distribution patterns identified for Japanese and Chinese respectively.

The distribution patterns that we identified were able to retrieve many correct matches with the numeral-classifier combination and target noun phrase. However, there were some instances when the noun phrase identified was incomplete, as shown in (3) below.

(3)	三千七百	名	会员
	<u>sānqiāngībǎi</u>	<u>míng</u>	<u>huìyuán</u>

几	普通市民
<u>jǐ</u>	<u>pǔtōngshìmín</u>

and citizen
“3700 members of the party and citizens.”

In (3), the target noun phrase picked out by the regular expression was “members of the party”. However, the entire target noun phrase should be “members of the party and citizens”. In such instances, we had to redefine the correct boundaries for the full noun phrase.

After retrieving the matches made by the regular expression, we tagged the classifiers to the target noun phrases by hand and marked the boundaries for both classifier and target noun phrase. We also marked the type of relationship they had: sortal, mensural, event, group, anaphoric, non-classifier and other. In cases where the target noun phrase was present in the sentence but was syntactically distant from the numeral classifier, as in (4), the relationship was marked as anaphoric.

(4)	苦情	は	毎月
	<i>kujou_T</i>	<i>wa</i>	<i>mai-getsu</i>
	complaints	<i>wa</i>	every-month
	平均 600件	に	上の
	<i>heikin</i> 600-ken ₁	<i>ni</i>	<i>noboru</i> ,
	average 600-CL	<i>ni</i>	add up
	前年	より	約
	<i>zen-nen</i>	<i>yori</i>	<i>yaku</i>
	last-year	than	about
	2000件	増	の
	2000-ken	<i>zou</i>	<i>no</i>
	2000-CL	increase	no
	18320件	を	摘発した
	18320-ken ₂	<i>wo</i>	<i>tekihatsushi-ta</i>
	18,320-CL	<i>wo</i>	expose-PST

“The number of complaints is as many as 600 per month on average and the police wrote tickets for 18,320 cases this year, up about 2,000 from last year.”

In the example, the target of the first 件 *ken*, a classifier used for things like “cases” or

¹ There can only be one classifier in an expression.

² DET: Determiner, NUM: Numeral, CL: Classifier, ?: 0 or 1, +: 1 or more

“matters”, is 苦情 *kujou* “complaint”. Therefore, the classifier was tagged to 苦情 *kujou* “complaint” with a sortal relationship. However, for the second *ken*, although the referent 苦情 *kujou* “complaint” is still in the sentence, it is not in the same clause as the classifier, therefore, for this classifier, it was tagged as anaphoric.

Anaphoric targets share a sortal relationship with the classifier that modifies the noun phrase. When the target is anaphoric, and the relationship between the target and classifier is under ‘other’, the set will be tagged as other instead of anaphoric.

Any instance of synecdoche was tagged as ‘other’. One instance of synecdoche was found for 名 *ming* (one of the classifiers for “people”), as shown in (5) below:

- (5) 六 名 自民党
liù *ming* *zìmǐndǎng*
 6 CL Liberal Democratic Party
 “6 members of the Liberal Democratic Party”

In this example, the numeral-classifier pair counts the number of party members and not the number of political parties.

Numerical-classifier combinations that were being used in an ordinal sense were tagged as ‘not’. In addition, noun phrases with very abstract referents like hope and courage were also tagged as ‘not’ as they were considered uncountable.

3.3 Assignment of synsets to classifiers

Then, we associated synsets from Japanese and Chinese Wordnet to each of these classifiers by hand. We looked up the semantic class for each target noun phrase and checked if it was suitable for the classifier. We also checked how high in the semantic hierarchy the use of the classifier could extend to. Table 3 illustrates how we assigned synsets to 个 *gè* (general classifier) and 只 *zhī* (animal classifier) based on the semantic hierarchy shown in Figure 1.

Classifier	Usage	Synsets
个 <i>gè</i>	general	+00001740-n -00015388-n
只 <i>zhī</i>	animal	+00015388-n -02374249-n -02512053-n -01726692-n
条 <i>tiáo</i>	fish	+02512053-n
	snake	+01726692-n
匹 <i>pǐ</i>	equine	+02374149-n

Table 3. Assignment of synsets to classifiers

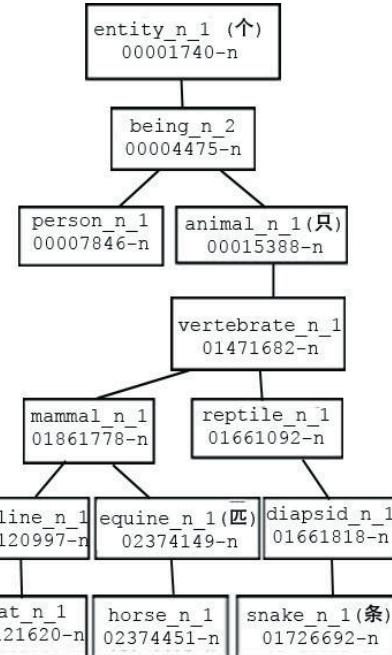


Figure 1. Semantic hierarchy in Chinese Wordnet

In both Chinese and Japanese, there is a general classifier that can be used for any entity when there is no specific classifier (个 *gè* for Chinese and 個 *ko* for Japanese). While 个 *gè* in Chinese can be used to count both “humans” and “objects”, 個 *ko* in Japanese cannot be used for “humans”.

Since 个 *gè* is considered a general classifier, we assigned the semantic class of entity_n_1 to it. A + symbol added in front of the synset signifies that all synsets below entity_n_1 will share the same classifier, 个 *gè*. We also removed animal_n_1 from it because animals are usually not counted with this classifier. The - symbol added in front of the synset signifies that the synset animal_n_1 does not share the same classifier.

只 *zhī* is the classifier used to count “animal”. We assigned animal_n_1 to it. We also removed fish_n_1, snake_n_1, equine_n_1 from it. These animals are not counted with the default animal classifier in Chinese. “Fish” and “snakes” are counted with 条 *tiáo*.

tiáo and “horses” and “mules” are counted with 匹 *pí*. There were quite a number of nouns that did not have an entry in both the Chinese and Japanese Wordnets. Thus, we had to add these nouns into the Wordnets. Due to time constraints, we only added nouns which occurred very frequently in the extracted noun phrases.

3.4 Generation of classifiers

We ran the annotated data through a program such that it picked out the head noun from a noun phrase (6). This is achieved by having the program go through the noun phrase and match it to a synset in the Wordnet, with the assumption that noun phrases are right-headed.

Based on example (6), the head noun that is retrieved from the noun phrase is 才才力ミ *ookami* “wolf”. The Chinese example in (7) shows that the head noun extracted and matched with a synset is 飞机 *fēijī* “aircraft”.

(6)	灰色	才才力ミ
	<i>hai-iyo</i>	<i>ookami</i>
	gray-colour	wolf
	“gray wolf”	
(7)	轻型	飞机
	<i>qīngxíng</i>	<i>fēijī</i>
	light	aircraft
	“light aircraft”	

We used the modified algorithm by Bond and Paik (2000) to generate the classifier. Based on this algorithm, the classifier most closely associated to the head noun in terms of semantic class was generated (Figure 2). All hyponyms under that synset will be counted with the same classifier unless a specific classifier has been marked for it.

Based on intuition, the algorithm will select the classifier marked on the closest possible hypernym. For example, the synset *antelope_n_1* is compatible with 只 *zhǐ*, which is the classifier for “animals” and 头 *tóu*, which is the classifier for certain types of animals like “pigs”, “cattle”, “elephants” and “livestock”. Since 头 *tóu* is associated with *bovid_n_1*, which is the hypernym of *antelope_n_1*, we select 头 *tóu* as the classifier for *antelope_n_1*.

-
- For a noun phrase:
- find its synset
 - find all hypernyms and see if any are selected for by classifiers
 - select the classifier from the synset with the highest similarity

Figure 2. Algorithm to generate a classifier

To analyse noun phrases with more than one sense, we used an algorithm shown in Figure 3. This algorithm identifies the semantic class of the noun based on the classifier used.

From the Japanese Wordnet, シイル *shiiru* “seal” has at least two senses:

- seal_n_9* marine mammal (subset of mammal)
- seal_n_5* a stamp affixed to a document (subset of stamp)

Since the classifier 匹 *hiki* is tagged to the semantic class of animal, the sense of シイル *shiiru* “seal” implied in the example is that of the synset *seal_n_9*, a marine mammal which is a hyponym of animal.

-
- For a noun phrase:
- Identify the semantic class of the head noun based on the classifier present

e.g. シイル 2 匹 買いました
shiiru 2-hiki kaimashita
 seal 2-CL buy-PST
 “I have two seals”

Figure 3. Algorithm to analyse a noun phrase

In this generation stage, we only looked at nouns with unique synsets that had been identified in the analysis stage. Lastly, we tested the predictions made using Wordnet to check the accuracy of the predictions.

4 Results

Japanese Classifier	Usage	Count
人 <i>nin</i>	“person”	122
件 <i>ken</i>	“(abstract) matters/cases”	69
台 <i>dai</i>	“vehicles”; “machines”	58
社 <i>sha</i>	“companies”; “shrines”	47

本 <i>hon</i>	long, thin objects e.g. “roads/ties/pencils”	45
枚 <i>mai</i>	thin, flat objects e.g. “pa- pers/photographs/plate s”	40
個 <i>ko</i>	general measure word; “military units”	30
点 <i>ten</i>	“pieces of a set”; “goods/items”	21
棟 <i>tou</i>	“buildings/apartments”	15
戸 <i>ko</i>	“houses”	15

Table 4. Ten most frequent Japanese classifiers

Chinese Classifier	Usage	Count
家 <i>jīā</i>	“families/businesses”	112
名 <i>míng</i>	“people”	107
个 <i>gè</i>	“people/objects”	111
场 <i>chǎng</i>	“events e.g. ex- ams/sporting events”	95
件 <i>jiàn</i>	“things/clothes”	93
位 <i>wèi</i>	“people” (honorific)	87
条 <i>tiáo</i>	“long and thin things e.g. snakes/rivers/ropes” “lives” “vehicles”	82
辆 <i>liàng</i>	“flat objects/things with flat surfaces e.g. beds, paper” “votes”	80
张 <i>zhāng</i>	“phrases/lines of verse/sayings”	74
句 <i>jù</i>		70

Table 5. Ten most frequent Chinese classifiers

Tables 4 and 5 show the ten most frequent classifiers in Japanese and Chinese respectively.

Classifier type	Japanese (J)	Chinese (C)
Sortal	592	1906
Anaphoric	133	113
Event	61	26
Group	7	41
Other	407	267
Non-classifier	142	921
Total	1400	3274

Table 6. Results of hand annotation

Table 6 summarises the results of the hand annotation for Japanese and Chinese. The total number of classifier phrases in Chinese is much more than in Japanese. This is because Chinese classifiers can also appear with determinatives like 这 *zhè* “this”, not just with numerals.

As shown in Table 6, the number of sortal classifiers in Japanese that share a sortal relationship with the target is less than half. This is much lower than we expected.

For Chinese, there were 1,906 noun phrases modified by a sortal classifier. This was slightly more than half of the total number of extracted sentences. Classifiers tagged as ‘other’ and ‘not’ were mostly being used in an ordinal sense or had uncountable abstract noun referents.

Scores	% (J)	Total (J)	% (C)	Total (C)
Correctly analysed	76.33	405	79.37	1312
Total	100	528	100	1653
Correctly generated	89.84	116	78.80	223
Total	100	129	100	283

Table 7. Analysis and generation scores

Table 7 presents the results of our evaluation.

The analysis score tells us how often we match a noun’s semantic class and the generation score tells us how often we correctly generate a classifier. A generated classifier is judged to be correct if it exactly matched the original classifier used in the annotated corpus. For Japanese, the analysis score is 76.33% and the generation score is higher at 89.84%. The analysis score is slightly lower by 3.04% than that for Chinese. However, the generation score is higher by 11.04% as compared to Chinese.

5 Discussion

For this study, we only considered classifiers that share a sortal relationship with the noun phrase they modify. Noun phrases that are modified by classifiers that share a group, event, mensural relationship were not included in the evaluation. Similarly, noun phrases in which the target is anaphoric were also not included.

Based on the results of the hand annotation, most sortal classifiers often have an anaphoric use. The anaphoric target can either be in the

same sentence but different clause or in a different sentence.

Overall, the evaluation of both algorithms is satisfactory. For Chinese, we were able to analyse correctly 79.37% or 1312 noun phrases. By using the default classifier assigned to each semantic class, we were able to generate correctly 78.80% or 223 classifiers.

For Japanese, we were able to correctly analyse 76.33% or 405 noun phrases and generate 89.84% or 108 classifiers.

One of the issues we faced in this study is the problem of dealing with synecdoche, particularly for Chinese. In example (5), given in Section 3.2, we saw the classifier 名 *míng* being used to count the number of members of the political party and not the number of political parties.

Based on our mapping of classifiers to semantic class, the “Liberal Democratic Party” would belong to the semantic class of organization which uses a different classifier 个 *gè*. However, this type of synecdoche will not be captured using the current method of analysis.

For Japanese, a large number of classifiers that were tagged with having “other” relationship with the targets were in fact functioning as ordinal classifiers. These were often preceded by an ordinal prefix (8) or followed by the ordinal suffix 目 *me* (9).

(8) 第一棟 の 旅館

dai-1-dou no ryokan
ORD-1-CL no hotel
“the first hotel”

(9) 二回目 の 優勝

ni-kai-me no yuushou
2-CL-ORD no victory
“the second victory”

One of the limitations of this study is the coverage of Wordnet. During our assignment of synsets to classifiers, we found that approximately 20% of our target noun phrases were not represented in CWN. For instance, 球队 *qiú duì* “a team for ball sports” was not included. Although it had synsets for 棒球队 *bàngqiúduì* “baseball team” (*baseball_team_n_1*) and 篮球队 *lánqiúduì* “basketball team”, (*basketball_team_n_1*), it did not have a generic term to refer to a team that played ball sports. For the Japanese study too, there are some nouns that are not yet represented in JWN, for example 大手 *oote* “major company” or チッショウ *tis-*

shuu “tissues”. In addition, the lexicon does not include proper nouns like names of companies like 三菱電機 *mitsubishi denki* “Mitsubishi Electric Corporation”. In a similar manner, a noun may be present in Wordnet but is missing the correct sense. One example is 白紙 *hakushi* “blank paper”. Although this noun is represented in JWN, the sense given is that of “fresh start”. The lack of a corresponding noun or sense in the lexicon may have affected the evaluation scores.

Similarly, some nouns which were represented in CWN were also missing other senses. For instance, 车 *chē* has two senses. The first sense is “car” and the second sense is “rook”, a type of chess piece. When we looked up 车 *chē* “car” CWN presented us with *rook_n_2*. This sense is the subset of corvine bird.

In this case, the noun is being represented by the wrong synset and one of its sense “car” was also missing. Although we assigned the synset for *vehicle_n_1* to the classifier 车 *liàng* (the classifier for “vehicles”), it was not able to generate this classifier when it encountered 车 *chē* “car” in the test sentences.

In addition, since the Chinese sentences were translated from Japanese sentences from the Mainichi Newspaper, there were also a few Japanese loanwords which were not represented in Chinese Wordnet. Some of these include 榻榻米 *tatāmi* “tatami” and 横岗 *hénggāng* “yokozuna”. This could also account for the lower generation scores obtained for Chinese as compared to Japanese. Although CWN covers more synsets, its coverage of common senses appears to be slightly worse than JWN.

Some nouns can also be used with more than one classifier. For instance, in Japanese, 住宅 *juutaku* “residence” can be used with classifiers 軒 *ken*, 個 *ko*, or 棟 *tou*, all three being classifiers for houses. In the corpus, there were instances of all three classifiers being used to quantify 住宅 *juutaku* “residence”. The choice of which classifier to use is up to the individual’s personal preference. Hence, it is difficult to predict the correct classifier for cases like this.

Another issue that may have contributed to the generation error is the problem of fixed expressions, particularly for Chinese. This was often seen with the classifier 口 *kǒu*, which is used for counting things with mouths. It is commonly

used to count the number of people in a family or household, as shown in (10).

(10) 一	家	五	口	人
	<i>yījiā</i>	<i>wǔ</i>	<i>kǒu</i>	<i>rén</i>
	a family	5	CL	person
	“a family of five.”			

Although (10) shows that 口 *kǒu* can be used to count “person”, this classifier is generally used this way only when it follows 家 *jiā* “family”. 个 *gè* is the more common classifier to use when counting “person”. Such fixed expressions cannot be properly analysed with our current methods of analysis and will require special processing.

Shape, size and animacy are some factors that play a part in selecting the correct classifier (Allan, 1977). For instance, 张 *zhāng* is used to count flat objects or things with a flat surface. Some examples include tables, stools, papers, newspapers and beds. However, the wordnets do not contain such information about shape or size of the nouns. Hence, some world knowledge is still required in order to predict the right classifier for a target noun phrase.

In order to further research on Japanese and Chinese classifiers, we release the following data for both languages under the creative-common attribution license (CC-by): (i) the table of classifier phrase + antecedent noun phrase pairs with their disambiguated synset (ii) the table of which synsets are classified by which classifiers (Figure 1). It is available from the Japanese Wordnet page: <http://nlpwww.nict.go.jp/wn-jp/>. We are also feeding back information on missing senses to the respective Wordnet projects.

6 Conclusion

In this paper, we presented an algorithm to generate numeral classifiers based on semantic hierarchies present in wordnets. For Chinese, it was shown to select the correct sortal classifier 78.80% of the time. We believe that this score can be raised with improvements to Chinese Wordnet. For Japanese, it was shown to select the correct sortal classifier 89.84% of the time. At the present moment, the wordnets do not provide a full coverage of all the nouns in the world. In addition, there are factors that may guide the choice of selection, making a purely taxonomic hierarchy inadequate. This study has shown that the selection of a classifier based only on a taxonomic hierarchy may not be accurate all the time because semantic attributes of the noun are also

important. Future studies can work to improve on the coverage of wordnet and also perhaps expand the wordnet in terms of linking semantic attributes. World knowledge is also required in order to select the most suitable classifier.

Acknowledgements

We wish to acknowledge the funding support for this project from Nanyang Technological University under the Undergraduate Research Experience on Campus (URECA) programme. We thank NICT for granting permission to use the NICT multi-lingual corpus.

References

- Allan, K. Classifiers. *Language*, 53:285-311, 1977.
- Bond, F. and Paik, K. Re-using an ontology to generate numeral classifiers. In *18th International Conference on Computational Linguistics: COLING-2000*, pp. 90-96, Saarbrucken, 2000.
- Downing, P. *Numerical classifier systems: The case of Japanese*. John Benjamins, Philadelphia: 1996.
- Fellbaum C. *WordNet*, MIT Press, 1998
- Guo, H., and Zhong, H. Chinese classifier assignment using SVMs. In *4th SIGHAN Workshop on Chinese Language Processing*, pp. 25-31, Jeju Island, 2005.
- Hundius, H. and Kölver, U. Syntax and semantics of numeral classifiers in Thai. *Studies in Language*, 7(2), pp. 165-214, 1983.
- Isahara H., Bond F., Uchimoto K., Utiyama M. and Kanzaki K. Development of Japanese WordNet. In *LREC-2008*, Marrakech, 2008
- Paik, K. and Bond, F. Multilingual generation of numeral classifiers using a common ontology. In *19th International Conference on Computer Processing of Oriental Languages: ICCPOL-2001*, pp. 141-147, Seoul, 2001.
- Sornlertlamvanich, V., Pantachat, W. and Meknavin, S. Classifier assignment by corpus based approach. In *15th International Conference on Computational Linguistics: COLING-94*, pp. 556-561, Kyoto, August 1994.
- Xu, R., Gao, Z., Pan, Y., Qu, Y., and Huang, Z. An Integrated Approach for Automatic Construction of Bilingual Chinese-English WordNet. In *Proceedings of ASWC2008*, pp. 302-314, Bangkok, Thailand, 2008.
- Zhang, H. Numerical classifiers in Mandarin Chinese. *Journal of East Asian Linguistics*, 16:43-59, 2007.
- Zhang, Y., Uchimoto, K., Ma, Q. and Isahara, H. Building an Annotated Japanese-Chinese Parallel Corpus – A Part of NICT Multilingual Corpora. In *10th Machine Translation Summit Proceedings*, pp. 71-78, Phuket, 2005.

Mapping a corpus-induced ontology of action verbs on ItalWordNet

Massimo Moneglia

University of Florence, Italy

moneglia@unifi.it

Alessandro Panunzi

University of Florence, Italy

alessandro.panunzi@unifi.it

Gloria Gagliardi

University of Florence, Italy

gloria.gagliardi@gmail.com

Monica Monachini

ILC CNR Pisa, Italy

monica.monachini@ilc.cnr.it

Francesca Frontini

ILC CNR Pisa, Italy

francesca.frontini@ilc.cnr.it

Irene Russo

ILC CNR Pisa, Italy

irene.russo@ilc.cnr.it

Abstract

Action verbs are the least predictable linguistic type for bilingual dictionaries and they cause major problems for NLP technologies. This is not only because of language specific phraseology, but it is rather a consequence of the peculiar way each language categorizes events. In ordinary languages the most frequent action verbs are “general”, since they extend productively to actions belonging to different ontological types. Moreover, each language categorizes actions in its own way and therefore the cross-linguistic reference to everyday activities is puzzling. A cross-linguistic stable ontology of actions is difficult to achieve because our knowledge on the actual variation of verbs across types of actions is largely unknown. This paper briefly presents the problems and the building strategies of the IMAGACT Ontology, which aims at filling this gap, and compares some early results on a set of Italian verbs with the information contained in ItalWordNet.

1 The Semantic Variation of Action verbs within and across Languages

Action verbs refer to at least an eventuality where an agent/causer theta role fills its argument structure. In all language modalities, action verbs bear the basic information that should be processed in order to make sense of a sentence. Especially in speech, they are the most frequent structuring elements (Moneglia and Panunzi, 2007), but unfortunately no one-to-one corres-

pondence can be established between an action verb, conceived as a lexical entry, and an action type, conceived as an ontological entity.

For instance, the English verb *to take* can refer to qualitatively different actions. In some uses the agent assumes the control of an object and changes its location (1); in some other uses the agent receives the object (2); in other cases, the agent takes the object away from somebody else (3):

- (1) John takes the umbrella
[= *to get*]
- (2) John takes the book from Sara
[= *to receive*]
- (3) the thief takes the money from the girl
[= *to take away*]

In short, in the above circumstances different action types occur. This judgment is confirmed by the productivity of each action type. For instance, despite the fact that the predicate is applied to different objects, humans are able to judge by reading a set of sentences whether the same action is performed or not:

- (1a) John takes the glass/ the candle
- (2a) John takes the pen from the assistant
- (3a) the thief takes the hat off the lady

Moreover, *to take* has several meanings corresponding to the different action types, and none of these types can be considered more appropriate than the others in characterizing the meaning of the verb. Each one could be a prototypic instance of the verb (Givon, 1986).

We call general verbs all natural language action verbs that share this property. In the case of general verbs, ordinary language does not mirror the ontology of action, causing a huge problem for all natural language understanding and machine translation tasks. As a matter of fact, the lemma does not specify the referred ontological entity.

2 Action Ontology and Translation

The problem of the lack of one-to-one mapping between lexical entries and ontological entities becomes even more relevant when cross-linguistic communication is taken into account.

The above variation of *to take* is also shared by the verbs roughly translating it in Italian. However, no translation equivalent can be established between action predicates of the two languages, as far as the ontological entity referred by action verbs is not identified and there is no guarantee that two predicates in a bilingual dictionary pick up the same entity (Majid et al., 2008). For this reason, action verbs are puzzling for machine translation, which may not find the translation equivalent even for simple sentences.

For example, according to pragmatic circumstances, the Italian sentence in (4) can be interpreted as an instance of different actions and can be translated into English respectively with *to take / to hold / to catch*, but this can be properly foreseen only if action types are identified cross-linguistically:

- (4) Mario prende il gatto
- (4a) Mario takes the cat
- (4b) Mario holds the cat
- (4c) Mario catches the cat

Given that action verbs have high frequency both in speech and in written corpora, this problem is extremely important for practical applications. The existence of the above semantic relations cannot be predicted, since they require general ontological knowledge which is not accessible through lexical entries.

Nevertheless, the mapping of general verbs to the action types is productive and should be in principle predictable. Once one action type is identified, we can foresee that the translation relation among predicates referring to that type in different languages holds in all instances of the type.

Therefore, action types can be considered as an ontological level that is independent from the language.

2.1 Action ontology and lexical databases: primary and marked variation

Existing verb typologies have gone a long way in systematically categorizing verbs into classes.

There is a range of lexical resources and ontologies which provide information on verb meaning variation (Baker et al., 1998; Levin, 1993; Kipper-Schuler, 2005; Palmer et al., 2005) and a number of initiatives which extend the information provided according to each frame for many languages.

The problems encountered by present ontologies to deal with categorization of action at cross-linguistic level can be made explicit by looking at WordNet (Fellbaum, 1998). Verbs are an important part of WordNet: more than 11,000 lexical entries, divided into 25,047 senses and corresponding to 13,767 synsets in the English database (version 3.0), which has been extended to many other languages.

For instance, WordNet identifies 42 synsets for the verb *to take*. Let's just focus on three of these entries:

- a) S: (v) take, get hold of (get into one's hands, take physically) "Take a cookie!"; "Can you take this bag, please";
- b) S: (v) lead, take, direct, conduct, guide (take somebody somewhere) "We lead him to our chief"; "can you take me to the main entrance?"; "He conducted us to the palace";
- c) S: (v) assume, acquire, adopt, take on, take (take on a certain form, attribute, or aspect) "His voice took on a sad tone"; "The story took a new turn"; "he adopted an air of superiority"; "She assumed strange manners"; "The gods assume human or animal form in these fables".

Despite its richness, this information is hard to use for disambiguation and translation tasks even by expert annotators (Ng et al., 1999). Since the glosses given for each synset are often too vague, the identification of the actual use of a verb among all its synsets becomes difficult.

Another crucial reason is that the productivity of verb application cannot be guaranteed by all synsets in the same manner. More specifically WordNet does not distinguish the synsets instantiating the proper meaning of the verb (for in-

stance a and b) from those which instantiate phraseological or metaphorical usages (for instance c).

Verbs have various usages which depart from their actual meaning, but those usages do not constitute any productive action type. From this perspective, it is reasonable to foresee that the Italian verb *prendere* can be applied to all instances of a (5) and in no instances of b (6):

- (5) he takes a cookie / a glass / a bag
- (5a) lui prende un biscotto / un bicchiere / una borsa
- (6) he takes the car / the dog / his friend there
- (6a) *lui prende la macchina / il cane / il suo amico là

On the contrary this is not the case of c (7-8), which is a metaphorical usage of the verb. We cannot foresee any regularity in the application of the Italian verb *prendere* to the possible instances of c:

- (7) he took an air of superiority
- (7a) ha preso un'aria di superiorità
- (8) he took on strange manners
- (8a) *ha preso strane maniere

In summary, despite the high number of senses registered in WordNet, there is no possibility of identifying those types that constitute the basis for a productive cross-linguistic relation.

3 The IMAGACT project

The IMAGACT project, which has been funded in Italy with the PAR/FAS program (undertaken by the University of Florence, ILC-CNR, Pisa, and the University of Siena) uses both corpus-based and competence-based methodologies for simultaneous extraction of a language independent action inventory from spontaneous speech resources of different languages.

The IMAGACT infrastructure faces key issues in ontology building. It grounds productive translation relations since it distinguishes the proper usage of verbs from their metaphorical or phraseological extensions; it allows easy identification of types in the variation, it is cross-linguistic in nature, it derives from the actual use of language but it can be freely extended to other languages through competence-based judgments

and it is therefore suitable for filling gaps in lexical resources.

3.1 Exploiting spontaneous speech corpora

The first idea developed in IMAGACT is to strictly define the relevant domain of language usage from which data about linguistic reference to actions can be derived. Actions specified by those verbs are most frequently used in ordinary communication since they are very relevant in everyday life. The actual use of action oriented verbs in linguistic performance can therefore be appreciated by observing their occurrence in spontaneous speech corpora, in which reference to action performance is primary.

The IMAGACT database focuses on high frequency verbs, which can provide sufficient variation in spoken corpora; i.e. roughly 500 verbs referring to actions which represent the full basic action oriented verbal lexicon.

In order to maximize the probability of occurrence of relevant action types, IMAGACT identifies the variation of this set in parallel on two spoken corpora:

- a 2 million word English corpus, taken from the British National Corpus;
- a collection of spoken Italian corpora with 1.6 million words in total (LABLITA corpus, Cresti and Moneglia, 2005; LIP, De Mauro et al., 1993; CLIPS corpus).

The corpus-based strategy consists of a manual annotation of the instances of action verbs, which first separates the metaphorical and phraseological usages from proper occurrences and then classifies proper occurrences into action types.

3.2 The cross-linguistic definition of the ontology of action in a Wittgenstein-like scenario

The experience in ontology building has shown that the level of consensus that can be reached in defining entities which are object of language reference tends to be lower, since the identification of such entities relies on a definition. Definitions are highly underdetermined, since they depend on the granularity of feature retrieval.

The traditional methodology will require reconciling in a unique definition all definitions given by linguists to classify the actions occurring in each language corpus. This is certainly a difficult task.

The key innovation of IMAGACT is to provide an alternative methodology which exploits the language independent ability to recognize similarities among scenes, distinguishing the identification of action types from their lexicographic definition. The annotator is required to identify the action that is expressed in each instance of the action verbs taken into account. Different actions are identified for each verb and then grouped into cross verbal action types.

All primary occurrences (the ones referring to a physical action) of a verb in the corpus are manually clustered around best examples that represent the different actions in terms of involved body schema, spatial relations and focal properties. Local equivalents (other verbs that can be replaced in the instance) are identified for each cluster, thus producing verb independent action types. Action types are then described in scripts and videos are made, onto which action types from many languages are mapped.

Working with more than one language produces a language independent type inventory. Crucially only the identification (and not the active writing of a definition) is required to set up the cross-linguistic relations. In Wittgenstein's terms, how can you explain to somebody what a play is? Just point out a play and say "this and similar things are plays" (Wittgenstein, 1953).

In IMAGACT the ontology makes use of the universal language of images which allows reconciling in a unique inventory of action types the descriptions derived from the annotation of corpora belonging to different languages.

For instance, let us consider the Italian verb *spingere* and the English verb *to push*, which might be expected to match on a similar set of action types. The annotation of the Italian corpus has identified six different types extended by *spingere*, which are instantiated by the following best examples:

- (9) il dottore spinge sulla pancia del paziente
[the doctor presses on the patient's belly]
- (10) John spinge la carta nel cestino
[John pushes the paper in the basket down]
- (11) la ragazza spinge il carrello
[the girl pushes the trolley]
- (12) Maria spinge la bottiglia giù dal tavolo
[Mary pushes the bottle off the table]
- (13) la ragazza spinge la creta nello stampo
[the girl pushes the plasticine into the mold]
- (14) lo yogi spinge il ventre avanti

[the yogi pushes his belly out]

On the basis of this information, a scene representing the occurrence of each type is produced. Therefore the above best examples will be respectively linked to A1, A5, A6, A8 and A9 of Figure 1 below.

Assuming that the English corpus will also be processed, the action types extended by *to push* will come about through best examples extracted from the corpus, obviously referring to different eventualities.

Of course there is no necessity that all possible types extended by *to push* and *spingere* will be recorded in the corpora, however the intersection of types actually extended by both verbs can be easily recognized. Indeed on the basis of the evidence provided by the cited scenes all competent speakers will recognize that, for instance, the Best examples "Mary pushes the car on the highway" should be mapped onto A5 and "The killer pushes the man off the cliff" should be mapped onto A6. This will be achieved without any direct comparison between Italian and English.

On the basis of the scenes, the differential of the two verbs for what regards their possible extension across action types recorded in the corpus will also easily be recognized. For instance, competent speakers will also recognize that A7 is the only model of Figure 1 in which *to push* cannot be extended at all. Considering the difference in extension between *spingere* and *to push* it will become evident that *to push* is more general since it can also be extended to A3, which is not a possible model of *spingere*.

IMAGACT will deliver a database of action types with their language encoding of English and Italian verbs in conjunction with the set of sentences (derived from corpora) instantiating each type.

Action types will be recorded in the form of videos. The scene corresponding to the best example of each action type (prototypic scene) is played by a supervisor. The adequacy of this scene in representing what's specified in the annotation is negotiated by the supervisor with the annotator to avoid misunderstandings. The scene is then transformed into a 3D animation and all information that is not essential to the representation is eliminated (stereotypic scene, not available at this stage of the project).

On the basis of this outcome it will be possible to ask informants with a different language what verb(s) is applied in their language for each type

identified by a scene and by a set of English sentences derived from corpus occurrences and assigned to that scene. The informant will provide the lexical choice available in his language. Crucially, the informant will verify whether or not the choice is correct for all arguments retrieved from corpus and assigned to that type.

4 Linking to WordNet

The IMAGACT project has already produced a corpus-based extraction of action types from a subset of high frequency Italian action verbs.

Let's consider as an example a restricted set of Italian verbs roughly equivalent in some way or other to *to press* or *to push*: {*spingere*, *premere*, *schiacciare*, *pigiare*, *comprimere*, *spremere*, *pressare*}.

The IMAGACT methodology starts from the identification of the semantic (referential) variation of the verbs. Once this variation is identi-

fied, it is possible to list all the referred action types, and then to connect to each type the set of equivalent verbs that can be applied to it.

From the given set of verbs a tentative inventory of action types has been extracted (see Figure 1).

Of course, each of the verbs that have been taken into account can be used to express a subset of these actions. For instance the verb *spingere* (the more general in the examined cases) can be applied to the actions A1, A2, A3, A5, A6, A8 and A9, while the verb *spremere* can be applied only to action A3.

Conversely, all identified action types can be referred by a set of verbs: Action A1 can be expressed by *spingere*, *premere*, *pigiare*, while Action A3 can be expressed by *schiacciare*, *comprimere*, *spremere*.

ACTION TYPE	DEFINITION (and example)	EQUIVALENT VERBS
A1 	Continuing generic pressure, with the sole result that the object (or body part) is pressed (<i>the doctor palpates the abdomen</i>)	<i>premere</i> , <i>pigiare</i> , <i>spingere</i>
A2 	Pressure on only one side that brings about the reduction of the volume of the object (<i>Jane compresses the garbage</i>)	<i>premere</i> , <i>pigiare</i> , <i>schiacciare</i> , <i>pressare</i> , <i>comprimere</i> , <i>spingere</i>
A3 	Pressure on two or more sides that brings about a reduction of the volume of the object (<i>John squeezes the toothpaste tube</i>)	<i>schiacciare</i> , <i>spremere</i> , <i>comprimere</i>
A4 	Non continuos (brief) pressure (<i>Jane presses the button</i>)	<i>premere</i> , <i>pigiare</i> , <i>schiacciare</i> , <i>spingere</i>
A5 	Continuous pressure that accompanies the object in the transition (<i>Jane pushes the trolley</i>)	<i>spingere</i>
A6 	Impulse that distances the object from the agent (<i>Jane pushes away the bottle</i>)	<i>spingere</i>
A7 	Pressure that causes damage to the object (<i>John crushes the tomato</i>)	<i>schiacciare</i>
A8 	Pressure that inserts the object into something (<i>John pushes the plasticine into the mould</i>)	<i>premere</i> , <i>pigiare</i> , <i>spingere</i>
A9 	Internal pressure (<i>the Yogi pushes the stomach out</i>)	<i>spingere</i>

Figure 1: IMAGACT action types

Speculatively we expect to find synsets that match these groupings, thus in this case one corresponding to A1 *{spingere, premere, pigiare}*, one corresponding to A3, and so on.¹

Still in the perspective of a linking of the final IMAGACT action inventory to one or more WordNets, among which ItalWordNet (IWN, Roventini et al., 2003), we could imagine at least three kinds of links:

- **Perfect matching:** an IMAGACT type of action matches a synset;
- **IMAGACT action types enriches IWN:** one or more IMAGACT types of actions are subsumed by a synset;
- **IWN enriches IMAGACT action types:** an IMAGACT action type subsumes one or more synsets.

The possibility of an imperfect match can also be foreseen.

In order to carry out the linking, a set of basic heuristics have been defined and applied to better align corpus-induced action verb types and the IMAGACT action types with the lexical knowledge encoded in ItalWordNet.²

We don't expect a full alignment but as a first step we aim at maximizing corpus-induced generalizations with synsets. In the early stages of the project we want to make clear the gap between lexical entries and ontological types.

A first check on ItalWordNet performed by a human annotator shows partial but not total matching for our predictions. Synsets have been retrieved by entering IWN with one of the lemmas associated with each action, and checking for synsets containing these lemmas. When all possible synsets seemed to be too generic, a best match has been found with hyponyms, if present.

¹ Let it be stressed that the action inventory that we are describing here is still under development, and it still lacks the important contribution that may derive from the analysis of English action verbs.

² The ItalWordNet lexical database (henceforth IWN) was first developed in the framework of the EuroWordNet project and then enlarged and improved in the national project SI-TAL1. The theoretical model underlying this lexicon is based on the EuroWordNet lexical model which is, in its turn, inspired by the Princeton WordNet (Fellbaum, 1998).

Act.	Synset	Type of match
A1	(<i>pigiare</i> [1], <i>premere</i> [1], <i>spingere</i> [5]) “apply pressure to something”	hypernymic match (the synset captures the meaning of the action but is more generic, applies also to A4)
A2	(<i>appiattire</i> [1], <i>schiacciare</i> [1]) “flatten”	possible exact match
A2	(<i>comprimere</i> [1], <i>pressare</i> [2], <i>schiacciare</i> [2]) “compress”	hypernymic match (A3, A8)
A3	(<i>comprimere</i> [1], <i>pressare</i> [2], <i>schiacciare</i> [2]) “compress”	hypernymic match (A2, A8)
A3	(<i>spremere</i> [1]) “squeeze”	hyponymic match (applies only when the object is soft)
A4	(<i>pigiare</i> [1], <i>premere</i> [1], <i>spingere</i> [5]) “apply pressure to something”	hypernymic match (also A1)
A5	(<i>imprimere un movimento</i> [1], <i>spingere</i> [2]) “make something move, push”	possible exact match
A6	(<i>spingere</i> [1]) “push away”	Possible exact match
A7	(<i>schiacciare</i> [3], <i>spiaccicare</i> [1]) “crush”	possible exact match
A8	(<i>comprimere</i> [1], <i>pressare</i> [2], <i>schiacciare</i> [2])	hypernymic match (A2, A3)
A8	(<i>pressare</i> [1]) “press”	no match
A9	(<i>premere</i> [3])	no match

Table 1: IMAGACT to ItalWordNet linking

This table shows how the simple access to IWN via the lemmas of potential action verbs does not guarantee finding good matches for the action at hand. As foreseen, on the one hand we have some synset like (*pigiare* [1], *premere* [1], *spingere* [5]) and (*comprimere* [1], *pressare* [2], *schiacciare* [2]) that are less specific than the action. On the other hand, a synset like (*spremere* [1]) seems to be more specific in that it applies to a subset of the objects that are involved in A3. Finally, some lemmas point to IWN synsets that, after manual inspection, prove to be non-relevant for the action at hand because of the existence of marked/metaphorical meanings.

Some heuristics can be applied to fine tune the linking: one possibility would be to use words that the annotator has identified as local equivalents in order to try and disambiguate between synsets. Local equivalents are manually identified verbs that cover the same action type (a notion akin to but not equal to that of synonymy).

For example, uses like *premere* for the action type A2 (“Fabio preme la carta nel cestino / Fabio presses the paper in the bin”) are not directly represented in IWN but they are recoverable through the local equivalent verb *pressare* and *schiacciare* that have synsets matching that action type. This strategy is useful but not resolute, because in several cases it is not possible to recover missing synsets or to narrow too general synsets, simply by looking at local equivalents’ synsets. However, in this case study we found that this heuristic was effective for 2 IMAGACT action types (A2, A3) out of 9.

The possibility of an automatic alignment of IWN with the action types in IMAGACT can also be taken into account. A possible strategy could be creating a link between each action and each synset showing a certain amount of match with the set of verbs expressing that action.

A perfect match would be that each of the verbs related to the action appear in the synset. This being rarely the case, an algorithm could compute similarity. When the same synset is linked to more than one action the link could be automatically identified as a hypernymic link, that is, IWN has fewer distinctions than the IMAGACT ontology. Clearly though this strategy is not error free: it does not perform any check on whether the synset is referring to a phraseological usage and most crucially it does not work when there is just one verb associated with an action, and that verb appears in many synsets. Moreover, the possibility still exist that a synset

including a totally different set of verbs (non-generic verbs for instance) can be found in IWN that matches the action we want to link.

5 Conclusions and future work

In this paper we briefly show that action verbs are the less predictable linguistic type for bilingual dictionaries and they cause major problems for NLP technologies because no one-to-one correspondence can be established between an action verb and an action type.

The need for general ontological knowledge which is not accessible through lexical entries motivates the IMAGACT project. It will use both corpus-based and competence-based methodologies for simultaneous extraction of a language independent action ontology from spontaneous speech corpora for different languages.

Although the project was just started, several issues concerning the initial version of a stable ontology of actions are already evident from the case study presented in this paper. We didn’t expect a full alignment but we aim at maximizing as a first step corpus-induced generalizations with synsets. As a first result we have shown that actions that are different for human annotators are not always mirrored by equivalent entries in lexical resources.

The comparison between corpus-induced generalizations about action types and lexical information found in ItalWordNet gives rise to a set of heuristics (i.e. using the hierarchy of IWN, checking for local equivalents’ synsets) that can be useful in the near future for a cross-linguistic integration of action types performed on the basis of English WordNet, after which the annotation process for English will be finished.

Acknowledgments

The IMAGACT project is funded within the PAR FAS Program of Tuscan Region (Italy), Action 1.1.a.3.

References

- Baker, C.F., Fillmore, Ch.J. and Lowe, J.B. 1998. The Berkeley FrameNet project. In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998)*. Montreal, Canada
- British National Corpus*, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: <http://www.natcorp.ox.ac.uk/>

- CLIPS Corpus.* URL: <http://www.clips.unina.it>
- De Mauro T., Mancini F., Vedovelli M., Voghera M. 1993. *Lessico di frequenza dell'italiano parlato*. Milano: ETASLIBRI.
- Majid, A., Boster, J.S., Bowerman, M. 2008. The cross-linguistic categorization of everyday events: A study of cutting and breaking. In: *Cognition*, 109, pp. 235-250.
- Fellbaum, Ch. (Ed.) 1998. *WordNet: An Electronic Lexical Database*. Cambridge (MA): MIT Press.
- Fillmore, Ch. J. and Atkins, B.T.S. 1992. Towards a frame-based organization of the lexicon: the semantics of RISK and its neighbors. In: Lehrer A., Kittay E.F. (Eds). *Frames, Fields, and Contrasts*. New Jersey: Lawrence Erlbaum Associates.
- Givon, T. 1986. Prototypes: Between Plato and Wittgenstein. In: Craig C. (ed.) *Noun Classes and Categorization*. Amsterdam: Beniamin, pp. 77-102.
- Kipper-Schuler, K. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD. Thesis. Computer and Information Science Dept, University of Pennsylvania, Philadelphia (PA).
- Korzen, I. 2005. Endocentric and exocentric languages in translation. In: *Perspectives: Studies in translatology*, 13 (1), pp. 21-37.
- LABLITA Corpus of Spontaneous Spoken Italian.* URL: <http://lablita.dit.unifi.it/corpora/>
- Levin, B. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.
- Moneglia, M., Panunzi, A. 2007. Action Predicates and the Ontology of Action across Spoken Language Corpora. The Basic Issue of the SEMACT Project. In M. Alcántara, T. Declerck, *Proceeding of the International Workshop on the Semantic Representation of Spoken Language* (SRSL7). Salamanca: Universidad de Salamanca, pp.51-58
- Ng, H.T., Lim, C.Y. and Foo, S.K. (1999). A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation. In: *Proceedings of the ACL SIGLEX Workshop on Standardizing Lexical Resources (SIGLEX99)*. College Park (MD), pp. 9-13.
- Palmer, M., Gildea, D. and Kingsbury, P. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. In: *Computational Linguistics*, 31 (1), pp. 71-106.
- Roventini A., Alonge A., Bertagna F., Calzolari N., Girardi C., Magnini B., Marinelli R., Speranza M., Zampolli A. 2003. ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian, in 'Linguistica Computazionale', Istituti Editoriali e Poligrafici Internazionali, Pisa-Roma, ISSN 0392-6907.
- Talmy, L. 1985. Lexicalization patterns: Semantic structure in lexical form. In: Shopen T. (Ed.) *Language typology and syntactic description, Vol. III: Grammatical categories and the lexicon*. Cambridge (UK): Cambridge University Press.
- Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

Using a Bilingual Resource to Add Synonyms to a Wordnet: FinnWordNet and Wikipedia as an Example

Jyrki Niemi and Krister Lindén and Mirka Hyvärinen

Language Technology, Department of Modern Languages, University of Helsinki
Helsinki, Finland

{jyrki.niemi,krister.linden,mirka.hyvarinen}@helsinki.fi

Abstract

This paper presents a simple method for finding new synonym candidates for a bilingual wordnet by using another bilingual resource. Our goal is to add new synonyms to the existing synsets of the Finnish WordNet, which has direct word sense translation correspondences to the Princeton WordNet. For this task, we use Wikipedia and its links between the articles of the same topic in Finnish and English. One of the automatically extracted groups of synonyms yielded ca. 2,000 synonyms with 89 % accuracy.

1 Introduction

Even a large wordnet is never complete but should be open to extending. Besides adding completely new senses (synsets), new synonyms can be considered for existing synsets. In this paper, we present a simple method for finding new synonym candidates for existing synsets of a wordnet by using a bilingual resource. We wish to extend the Finnish wordnet, and we use Wikipedia as the source for new synonyms.

1.1 FinnWordNet as a Translation

FinnWordNet – The Finnish WordNet (FiWN)¹ was initially created by translating into Finnish all the word senses in Princeton WordNet (PWN, version 3.0) (Fellbaum, 1998). FiWN has 117,659 synsets. The first version of FiWN was published in December 2010; the current version with some corrections is 1.1.2. FiWN is freely available under the Creative Commons 3.0 licence (CC-BY).

The PWN word senses were translated by professional translators to ensure the quality of the content. The translation process is outlined and

discussed by Lindén and Carlson (2010). The direct translation approach was based on the assumption that most synsets in PWN represent language-independent real-world concepts. Thus also the semantic relations between synsets are assumed to be mostly independent of the language, so the structure of PWN can be reused as well. This approach made it possible to create an extensive Finnish wordnet directly aligned with PWN. The direct translation of PWN word senses from English into Finnish also provided us with a translation relation and thus a bilingual wordnet.

The work described in this paper also acts as an evaluation of the translations: the quality can be considered the better the fewer translations need to be corrected, and the coverage the better the fewer translations need to be added.

1.2 Extending FinnWordNet

We wish to extend FiWN in various, preferably semi-automatic ways. In this paper, we consider adding missing synonyms to existing synsets. For example, the synset containing the words *cover* and *blanket* lacks the common Finnish word *peitto*, although the existing translations are valid.

Adding thousands of words to their correct places in the semantic hierarchy is a tedious task if done manually. Hence our focus is on such words that can be automatically placed into the structure, i.e. Finnish words with an English translation found in PWN. Since the structure of FiWN follows that of PWN, we can assume that the Finnish equivalent of any English word sense in PWN belongs in the corresponding place in FiWN.

The rest of this paper is organized as follows. Section 2 reviews some related work. Section 3 describes the method for finding new synonym candidates. Section 4 describes the Wikipedia data used for evaluation; Section 5 presents evaluation results. Section 6 discusses the results and avenues for future work, and Section 7 concludes the paper.

¹<http://www.ling.helsinki.fi/en/lt/research/finnwordnet/>

2 Related Work

Wikipedia² and its sisters, such as Wiktionary,³ have been exploited in various NLP tasks and also as sources of lexical information. An extensive overview on such tasks as well as the structure of Wikipedia is presented by Medelyan et al. (2009).

Our approach resembles that of Tyers and Pienaar (2008), who extracted bilingual translation lexicons from the interlanguage links in Wikipedia. Erdmann et al. (2009) extracted domain-specific terminologies with a similar method, but since a Wikipedia article has at most one interlanguage link for each language, they also obtained synonymous translations from redirection pages and link texts. By contrast, we do not construct a dictionary from scratch but use the existing data in FiWN.

Alkhalifa and Rodríguez (2009) use the English–Arabic interlanguage links in Wikipedia to add new named entities (synsets) to the Arabic WordNet, corresponding to ones in PWN. By contrast, in this work, we only search for new synonyms for existing synsets.

3 Method for Finding Synonym Candidates in a Bilingual Resource

Our method essentially mines new synonyms for a wordnet by using translation pairs in a bilingual resource (BLR) aligned at word (or phrase) level, by joining them on the English word in PWN, and by considering the Finnish translations found in the BLR as synonym candidates for FiWN. The principle is illustrated in Fig. 1. The synonym candidates must then be manually checked for correctness before adding them to FiWN.

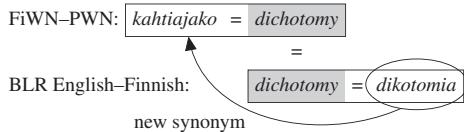


Figure 1: Finding a new Finnish synonym by joining on the English word: *dikotomia* as a synonym for *kahtiajako*, both translations of *dichotomy*.

We are able to apply the method to FiWN because it has been created by directly translating the word senses of PWN, which provides a translation relation between Finnish and English word senses. The PWN–FiWN translation relation is between

²<http://www.wikipedia.org>

³<http://www.wiktionary.org>

individual word senses in synsets instead of between synsets as in many other multilingual wordnets, such as EuroWordNet (Vossen, 1998). The translation relation is many-to-many.

When an English word is present in both PWN and the BLR, there are four different basic occurrence categories for the translation in FiWN (illustrated in Fig. 2):

1. FiWN already has the exact translation pair.
2. FiWN has the translation for a different word in the same synset.
3. FiWN has the translation for a different word in a different synset.
4. No synset in FiWN has the translation as a synonym.

We classify each translation pair only into the first matching category.

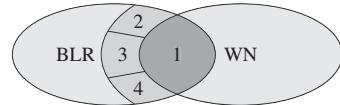


Figure 2: The categories of translation pairs from two sources: a bilingual resource and a wordnet.

The translation pairs in category 1 can be disregarded, since they already occur in PWN–FiWN. The Finnish words in the translation pairs in category 2 could in general be added to FiWN directly, without manual checking, if desired.

The translation pair categories in which we are mainly interested in this paper are 3 and 4, since in these categories the found translation pair does not occur in the PWN–FiWN translation relation. Each of these two categories is further divided into two different groups based on whether the English word occurs in (a) one or (b) more PWN synsets.

More formally, let WN be the PWN–FiWN translation relation where $(w_{en}^W, w_{fi}^W, ss) \in WN$, w_{en}^W is the English word in PWN, w_{fi}^W its Finnish translation in FiWN and $ss \in SS$ a synset identifier. Let BLR be the translation relation in the BLR, $(w_{en}^B, w_{fi}^B) \in BLR$. We then consider the join J between BLR and WN on the English word: $J = BLR \bowtie_{w_{en}^B=w_{en}^W} WN$, where $(w_{en}, w_{fi}^B, w_{fi}^W, ss) \in J$ and $w_{en} = w_{en}^B = w_{en}^W$. A translation pair $tp = (w_{en}, w_{fi}^B)$ is a projection of tuple $(w_{en}, w_{fi}^B, w_{fi}^W, ss) \in J$. The above categories can be defined for tp as follows:

1. For some $(w_{en}, \cdot, w'_{fi}^W, \cdot) \in J$: $w_{fi}^B = w'_{fi}^W$.
2. For some $(w'_{en}, w_{fi}^B, w'_{fi}^W, ss) \in J$: $w'_{en} \neq w_{en}$.
3. For some $(w'_{en}, w_{fi}^B, w'_{fi}^W, ss') \in J$: $w'_{en} \neq w_{en}$ and $ss' \neq ss$.
4. For all $(\cdot, \cdot, w'_{fi}^W, \cdot) \in J$: $w'_{fi}^W \neq w_{fi}^B$.

4 Test Data

4.1 Wikipedia Translation Links

To test our method outlined above, we used as the bilingual resource the interlanguage (translation) links between the Finnish and English Wikipedia in the freely available article contents of the Finnish Wikipedia as of 29 August 2011.⁴ From a Wikipedia article, we extracted its title and the interlanguage links containing the title of the target article prefixed with a language code.

4.2 Preprocessing and Filtering Translations

A Wikipedia article title may contain in parentheses a *disambiguation tag* disambiguating between different senses of a word. To simplify our task, we filtered out all article titles with a disambiguation tag, along with the titles of disambiguation pages, since they are by definition polysemous.

Because the titles of Wikipedia articles are in general nouns or noun phrases, we regarded only the nouns in FiWN when considering translations.

We included article titles with a namespace prefix; we simply removed the prefix. We omitted translation links pointing to a section of an article.

We considered the Finnish Wikipedia title to be equal to the FiWN word (and classified in category 1) even if they differed in capitalization. We also lemmatized the Finnish words in FiWN and the Finnish Wikipedia and considered the words (or phrases) equal if their lemmas were the same. As the lemmatizer we used Omorfi.⁵

4.3 Data Sizes

The number of Finnish–English noun translation pairs in the PWN–FiWN translation relation is 157,775 and those in the interlanguage links from the Finnish to English Wikipedia 213,796.

Table 1 shows the number of different Wikipedia article titles in Finnish and English, the number of different nouns (noun phrases) in FiWN and

⁴<http://download.wikimedia.org/fiwiki/20110829/fiwiki-20110829-pages-articles.xml.bz2>

⁵<https://gna.org/projects/omorfi>

PWN, and how many of them are in common.⁶ To make the numbers comparable with the number of translation pairs, we preprocessed and filtered the words as described in Sect. 4.2.⁷ The numbers for Wikipedia include the titles of various meta pages as well as articles proper.

	Finnish	English
Unique WP titles (WP)	326,546	5,543,618
Unique WN nouns (WN)	100,901	117,972
Common ($C = WP \cap WN$)	19,974	38,985
Common of WN (C / WN)	19.8 %	33.0 %

Table 1: Wikipedia titles in FiWN and PWN.

The Finnish Wikipedia interlanguage links contained 25,062 different translation pairs in which the English word was found in PWN. In preprocessing, we filtered out 8,148 of them, leaving us with 16,914 Finnish synonym candidates.

5 Results and Evaluation

5.1 Classifying Synonym Candidates

The translation pairs obtained from the join of the Wikipedia and wordnet were divided into the categories described in Sect. 3. In addition, we counted untranslated words, which were identical in the Finnish and English Wikipedia, FiWN and PWN, mostly proper nouns. The number of translation pairs in each category and their percentage of the total are listed in Table 2.

Category	Translation pairs	% of pairs
Untranslated	3,451	20.4
1	8,478	50.1
2	1,245	7.4
3a	554	3.3
3b	356	2.1
4a	2,278	13.5
4b	552	3.3
Total	16,914	100.0

Table 2: The number of translation pairs in each category and their percentage of the total.

The data sets 1 to 4 and the evaluated samples of data sets 3 and 4 are available for download.⁸

⁶The figures for the English Wikipedia are based on the dump of article contents on 4 August 2011.

⁷The intersection for English is smaller than the 80,295 reported by Navigli and Ponzetto (2010) because we have omitted the titles of disambiguation and redirection pages.

⁸<http://www.ling.helsinki.fi/en/ltr/research/finnwordnet/testdata/gwc2012/>

Synonym candidate quality	Category					Total
	3a	3b	4a	4b		
Replaces original translation	4 (3.8)	2 (2.7)	5 (2.2)	2 (1.8)	13 (2.5)	
To be added	68 (64.8)	29 (39.2)	198 (86.8)	65 (59.1)	360 (69.6)	
– good as such	62 (59.0)	26 (35.1)	145 (63.6)	53 (48.2)	286 (55.3)	
– good if edited	4 (3.8)	3 (4.1)	8 (3.5)	7 (6.4)	22 (4.3)	
– alternative form	2 (1.9)	0 (0.0)	45 (19.7)	5 (4.5)	52 (10.1)	
Unsure	1 (1.0)	1 (1.4)	3 (1.3)	0 (0.0)	5 (1.0)	
Poor	32 (30.5)	42 (56.8)	22 (9.6)	43 (39.1)	139 (26.9)	
Total (sample size)	105 (100.0)	74 (100.0)	228 (100.0)	110 (100.0)	517 (100.0)	

Table 3: The quality of the synonym candidates found in the samples of data categories 3 and 4. The numbers in parentheses are percentages.

5.2 Evaluating Synonym Candidates

We tested our method using ca. 20 % samples of the data, except for the largest set 4a, for which we deemed a 10 % sample to suffice for reliable enough results. As we focus on new synonym candidates, we do not analyse category 1, which contains only translation pairs already in FiWN.

The sample for category 2 was rather homogeneous, containing pairs in which the Finnish word is a good translation for the English one as such, but slightly less precise than the FiWN translation. For instance, *Amur-joki* was suggested as a translation for *Amur* in a synset containing the translation pairs *Amur = Amur* and *Amur River = Amur-joki* (*joki* means ‘river’). However, we decided against adding such less precise translations. Losing the fine distinctions between the translations of word senses would also move the PWN–FiWN translation relation towards one between synsets.

For categories 3 and 4, we determined the number of synonym candidates replacing the translation in FiWN, ones to be added, and poor and unsure ones. The synonym candidates to be added were further divided into good ones as such, good ones if edited (e.g., from plural to singular) and alternative forms of the translation in FiWN. Alternative forms included variants of dates and proper nouns as well as alternative spellings. Unsure candidates included medical terminology and complicated abstract concepts. These results are shown in Table 3. As can be seen, the translations missing altogether from FiWN (category 4) are much more often useful than the ones already occurring in some synset (category 3).

Based on the results from the samples, we estimated the total number of synonyms that could be mined from the Finnish Wikipedia interlanguage

link data with using our method. We used the percentages obtained from the samples for the whole data set. We considered separately synonyms that would replace the translation in FiWN and synonyms to be added (all subgroups together). Table 4 shows the estimated numbers by category, along with confidence estimates calculated based on the size of the sample of each category.

An example of a found new synonym is *peitto* for the synset containing the words *cover* and *blanket* in PWN. Although the existing synonyms *peite* and *huopa* are good translations of the English words, *peitto* is the most common Finnish word for a blanket as described in the synset gloss. We also found several official terms and loanwords; for instance, *eristic*, translated as *väitety*, was offered the additional translation *eristiikka*.

A poor synonym candidate is often a good translation of the English word, but in the wrong sense for the synset for which it is suggested. The different senses of a polysemous English word tend to be translated as several different words in Finnish.

6 Discussion and Future Work

All in all we found in Wikipedia 16,914 translation pairs which could be relevant for FiWN (version 1.1.2). Among the relevant synonym candidates in categories 3 and 4, we estimated that 91 of 16,914 (= 0.5 %) were to replace the original translation and 2,803 of 16,914 (= 16.6 %) were to be added. From this we can conclude that the quality of the original translations from PWN to FiWN is high.

The translation pair category that provided the best results was clearly 4a (89.0 ± 4.1 % useful synonym candidates). The synonyms provided by this group do not yet occur in FiWN and they are translations of English words monosemous in

Type	Category					Total
	3a	3b	4a	4b		
Replace	21±20 (3.8±3.7)	10±13 (2.7± 3.7)	50± 43 (2.2±1.9)	10±14 (1.8±2.5)	91± 50 (2.4±1.4)	
Add	359±51 (64.8±9.1)	140±40 (39.2±11.1)	1978±100 (86.8±4.4)	326±51 (59.1±9.2)	2803±148 (74.9±4.0)	
Total	380±49 (68.6±8.9)	149±40 (41.9±11.2)	2028± 92 (89.0±4.1)	336±50 (60.9±9.1)	2893±145 (77.4±3.9)	

Table 4: Estimated total number of replacement and additional synonyms for FiWN obtainable from the Finnish Wikipedia data. The numbers in parentheses are percentages of the total number of translation pairs in each category. Confidence estimates are at the 95 % confidence level.

PWN. This category was also by far the largest of the relevant ones: we estimated that it could yield roughly 2,000 new synonyms to FiWN.

If we knew with reasonable certainty which Wikipedia articles correspond to which synsets, we could improve the accuracy of in particular those synonym candidates which have several possible target synsets. Ruiz-Casado et al. (2005) present a method for linking Wikipedia articles and WordNet synsets based on the similarity between the content of the Wikipedia article and the gloss of the synset. Navigli and Ponzetto (2010) use WordNet synonyms, hypernyms, hyponyms and sister words, as well as glosses, in determining correspondences between WordNet synsets and Wikipedia articles. The method of Niemann and Gurevych (2011) allows multiple alignments between synsets and Wikipedia articles. Even if imperfect, such methods can speed up the manual verification by often providing good suggestions.

7 Conclusion

In this paper, we presented a method for finding new synonym candidates for synsets in the Finnish WordNet, which has direct word sense translation correspondences to the Princeton WordNet. The method exploits translation relations in bilingual resources having the same languages. We tested the method with the Finnish–English interlanguage links of the Finnish Wikipedia. Only $0.5 \pm 0.3\%$ of the suggested synonyms were estimated to replace a translation already in FiWN, which indicates a good quality of translation. The evaluation of a sample of the synonym candidates that do not occur in FiWN and that are translations of monosemous PWN words showed that we could add $89.0 \pm 4.1\%$ of such synonyms.

References

- Musa Alkhalifa and Horacio Rodríguez. 2009. Automatically extending NE coverage of Arabic WordNet using Wikipedia. In *Proceedings of 3rd International Conference on Arabic Language Processing (CITALA'09)*, pages 23–30, Rabat, Morocco, May.
- Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2009. Improving the extraction of bilingual terminology from Wikipedia. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5:31:1–31:17, November.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, May.
- Krister Lindén and Lauri Carlson. 2010. FinnWordNet – WordNet på finska via översättning. *LexicoNordica – Nordic Journal of Lexicography*, 17:119–140.
- Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from Wikipedia. *Int. J. Hum.-Comput. Stud.*, 67:716–754, September.
- Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010*, pages 216–225.
- Elisabeth Niemann and Iryna Gurevych. 2011. The people’s web meets linguistic knowledge: Automatic sense alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics*, pages 205–214, Oxford, UK, January.
- Maria Ruiz-Casado, Enrique Alfonsena, and Pablo Castells. 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In *Advances in Web Intelligence*, volume 3528 of *Lecture Notes in Computer Science*, pages 947–950. Springer, Berlin / Heidelberg.
- Francis M. Tyers and Jacques A. Pienaar. 2008. Extracting bilingual word pairs from Wikipedia. In *Collaboration: interoperability between people in the creation of language resources for less-resourced languages (A SALTMIL workshop)*, pages 19–22, May.
- Piek Vossen, editor. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Dordrecht.

Building WordNets by machine translation of sense tagged corpora*

Antoni Oliver, Salvador Climent

Universitat Oberta de Catalunya (UOC)
Avda. Tibidabo 39-43 08035 Barcelona
aoliverg,scliment@uoc.edu

Abstract

This paper describes a methodology for the construction of WordNets based on machine translation of an English sense tagged corpus. For the construction of such a corpus we used two freely available resources: the Semcor Corpus and the Princeton WordNet Gloss Corpus. This methodology is applied on the construction of Spanish and Catalan WordNet 3.0. In our first experiments we used a simple word alignment algorithm obtaining precision results comparable to those obtained by methods based on bilingual dictionaries. After that, we used a freely available statistical word alignment algorithm (Berkeley Aligner) obtaining better results. This methodology can be suitable for those languages with an available statistical machine translation system and can be used for either constructing WordNets from the scratch or for enlarging existing WordNets.

1 Introduction

WordNet (Fellbaum, 1998) is a lexical knowledge database that organizes nouns, verbs, adjectives and adverbs in sets of synsets. Each synset represents a lexicalized concept in English. Synsets are connected to other synsets by semantic relations (hiponymy, antonymy, meronymy, troponymy, etc.). Each synset has a gloss or definition. WordNet has become a standard resource in all kind of researches and applications in Natural Language Processing.

The English WordNet (PWN *Princeton WordNet* henceforth) is a free resource and it's available

*This research has been carried out thanks to the Project MICINN, TIN2009-14715-C04-04 of the Spanish Ministry of Science and Innovation

for download from its web page of the University of Princeton¹. The current version is 3.1 but at the moment of writing this paper the version available for downloading is 3.0.

Several projects have developed WordNets for other languages: EuroWordNet (Vossen, 1998) initially for Dutch, Italian, Spanish and for the enlargement and improvement of English, and in an extension of the project for German, French and Estonian; Balkanet (Tufis et al., 2004) for Bulgarian, Greek, Romanian, Serbian and Turkish and RusNet (Azarova et al., 2002) for Russian, among others. On the Global WordNet Association² website we can find a comprehensive list WordNets available for different languages.

Many of the available WordNets are subject to proprietary licenses. Some of the WordNets other than PWN holding a free license are: Catalan, Danish, French WOLF WordNet, Hebrew, Hindi, Japanese, Russian and Tamil WordNets among others. Our goal is to improve Catalan WordNet 3.0 and develop Spanish WordNet 3.0 and distribute both under free licenses.

2 WordNet building strategies

In this section we review some techniques that have been used to build WordNets for various languages other than English. We can distinguish two general approaches for building WordNets (Vossen, 1998):

- **Merge model:** a new ontology is constructed for the target language, with its own levels and relations. After that, relations between PWN and this local WordNet are generated.
- **Expand model:** English variants associated with PWN synsets are translated, using bilingual dictionaries or other strategies. Subsequently, relations imposed by the structure

¹<http://wordnet.princeton.edu>

²<http://www.globalwordnet.org>

of PWN are assumed to be valid for the target language. It is not necessary to establish interlingual relations because local WordNet and PWN are parallel.

Each of these strategies has a number of pros and cons (Vossen, 1996). On the one hand, the expand model is simpler from a technical point of view and ensures a greater degree of compatibility between WordNets for different languages. In contrast, WordNets developed using this technique are overly influenced by PWN and thus retain their mistakes and structural drawbacks. The merge model strategy is more complex from a technical point of view but allows a more direct use of existing ontologies and thesauri.

In the EuroWordNet project (Vossen, 1999) Dutch and Italian used the merge model, whereas Spanish chose the expand model. Catalan WordNet used the same methodology as Spanish WordNet (Benítez et al., 1998). Other projects using the merge model are RussNet (Azarova et al., 2002) and BalkaNet (Tufis et al., 2004). The expand model has been used as well in the MultiWordNet project for Italian (Pianta et al., 2002), the French WOLF WordNet (Sagot and Fišer, 2008), the Indonesian WordNet (Putra et al., 2008) and the Hungarian WordNet (Mihálitz et al., 2008), among others.

Some WordNets have been constructed using a hybrid approach. In the Portuguese WordNet (Marrafa, 2002), for instance the vocabulary and the set of internal relations were developed independently from PWN but PWN was taken as model and source. Basque WordNet (Agirre et al., 2002) used also a hybrid approach since PWN was taken as a starting point but new hierarchies and lexical-semantic relations extracted from dictionaries were also used to enrich, and in some cases, modify those of PWN.

3 Using machine translation in the construction of WordNets

In this section we will briefly present two projects related to WordNet that use machine translation systems to translate a semantically tagged corpus: the construction of the Macedonian WordNet and the Babelnet project.

3.1 The Macedonian WordNet

In the construction of the Macedonian WordNet (Saveski and Trajkovski, 2010) a bilingual

English-Macedonian dictionary has been used as a main source of information. For monosemic entries the PWN synset - Macedonian word relation can be established directly. For polysemic entries the task of establishing relations between PWN synsets and Macedonian words can be seen as a Word Sense Disambiguation problem. In this paper the authors used the proposal of (Dagan and Itai, 1994) based in the fact than a given word tends to co-occur more times in a big corpus with the open class words from its definition than with other words. The authors had to face two main problems: (i) the available English-Macedonian dictionary didn't have definitions; and (ii) no big Macedonian corpora was available. The first problem was solved by translating the PWN glosses into Macedonian using Google Translate; in this way, translated glosses were used as definitions. The second problem was solved using the web as a corpus through the *Google Similarity Distance* (Cilibrasi and Vitanyi, 2007). This distance between a word or a phrase x and a word or a phrase y is calculated using the Google search results for x , for y and form a query including x and y .

3.2 The Babelnet project

The main goal of the Babelnet project (Navigli and Ponzetto, 2010) is the creation of a big semantic network by linking the lexicographic knowledge from WordNet to the encyclopedic knowledge from Wikipedia. This is done by assigning WordNet synsets to Wikipedia entries. Since Wikipedia entries bear interlingual relations, a variant in several languages can be assigned to some of the WordNet synsets. Therefore, if a relation between a synset s and an English Wikipedia entry w_{eng} has been set, using the Wikipedia interlingual links the same relation can be set for all languages having the corresponding Wikipedia entry: w_{spa} , w_{fra} , etc. For those languages lacking the corresponding Wikipedia entry, the authors propose the use of Google Translate to translate a set of English sentences containing the synset s . This set of sentences is built using two sources:

- The Semcor corpus (Miller et al., 1993).
- Sentences from Wikipedia containing a link to the English Wikipedia page w_{eng} .

Once the automatic translation is done the most frequent translation is detected and included as a variant for the synset s in the given language.

4 Our approach

4.1 Goal

In this paper we propose an approach for the construction of WordNets based on machine translation of sense tagged corpora. As it is known, it is relatively easy to extract a bilingual dictionary from a parallel corpus (Och and Ney, 2003). If the sentences of the source language are semantically tagged with PWN synsets, we can treat these synsets as words so that we can get their corresponding variants for the equivalent synsets in the target language. To our knowledge it doesn't exist a big semantically tagged bilingual corpus for English and either Spanish or Catalan -or at least tagged for the English source. For this reason we use machine translation systems to get such a parallel corpus from an English monolingual corpus.

4.2 Sense tagged corpus

We have used two sense tagged corpora for English (see 4.6 below) where the tags are the PWN 3.0 synsets. For each sentence we can get two versions: the sentence itself, and the sentence with all tagged words substituted by its PWN 3.0 synset. We can see an example in the following lines:

Then he noticed that the dry wood of the wheels had swollen .
00117620r he 02154508v that the 02551380a
15098161n of the 04574999n had 00256507v .

4.3 Use of Machine Translation

The machine translation system to be used in this work must be capable to perform good lexical selection, that is, it should select the correct target words for the source English sentence. For ambiguous words the system must be able to disambiguate and choose the correct translation. Other translation errors are less important for our experiments. For our experiments we use two statistical machine translation systems: Google Translate and Microsoft Bing Translator. We didn't assess in deep the ability of these systems to do a correct lexical selection, but we performed some successful tests, as explained now on. Consider first sentences containing the English word *wood*. This word is a variant of both the PWN synset 15098161-n (*the hard fibrous lignified substance under the bark of trees*; translated into Spanish as *madera* and into Catalan as *fusta*) and 08438533-n (*the trees and other plants in a large densely wooded area*; Spanish *bosque* and Catalan *bosc*). If we take a sentence corresponding to the first sense

and we translate it using the given MT systems we get:

This house is made of wood.

*Google Translate:

Esta casa es de madera.

Aquesta casa és de fusta.

*Microsoft Bing Translator:

Esta casa està hecha de madera.

Aquesta casa és fa de fusta

By performing the same task to a sentence corresponding to the second sense we get:

He got lost in the wood beyond Seattle.

*Google Translate:

Se perdió en el bosque más allá de Seattle.

Es va perdre en el bosc més enllà de Seattle.

*Microsoft Bing Translator:

Se perdió la madera más allá de Seattle.

Es va perdre en la fusta més enllà de Seattle.

Consider now the English word *bank*. As a noun it has 10 meanings according PWN. We will concentrate on two of them: 09213565n (*sloping land (especially the slope beside a body of water)*) and 08420278n (*a financial institution that accepts deposits and channels the money into lending activities*). The first meaning has three possible variants in Spanish (*margen, orilla, vera*) and two in Catalan (*marge, vora*). The second meaning has one variant both in Spanish (*banco*) and Catalan (*banc*)³.

If we take a sentence corresponding to the first sense and we translate it with the given MT systems we get:

She waits on the bank of the river.

*Google Translate:

Ella espera en la orilla del río.

Ella espera a la vora del riu.

*Microsoft Bing Translator:

Ella espera en la orilla del río.

Ella espera a la riba del riu.

and if we perform the same task to a sentence corresponding to the second sense we get:

She puts money into the bank.

*Google Translate:

Ella pone el dinero en el banco.

Ella posa els diners al banc.

*Microsoft Bing Translator:

Ella pone dinero en el Banco.

Ella posa diners en el Banc.

As we can see, these systems can do, at least in certain situations, a good lexical selection (note that Microsoft translator failed to select the correct Spanish and Catalan translation for *wood* as a forest, using *madera* instead of *bosque* and *fusta* instead of *bosc*). In the second example we can observe an interesting result. While in Spanish both

³All these variants were taken from the preliminary versions of Spanish and Catalan WordNet 3.0.

systems propose a translation for bank (of a river) that coincides with one of the variants of Spanish WordNet, in Catalan Microsoft Translator proposes a translation (*riba*) not registered as a possible variant in the Catalan WordNet, but that can be considered correct. In some situations, the use of machine translation systems can be useful for enriching existing local WordNet versions.

The machine translation systems fail to choose the correct word in some cases, thus this will undoubtedly lead to errors. Nevertheless, our evaluation methodology, described in section 5, gives us an idea of the influence of the translation quality in the results.

4.4 Bilingual corpus

Using a machine translation system we can get a bilingual corpus. For example, from the sentence used in 4.2 above we can get the pair English - Spanish:

Then he noticed that the dry wood of the wheels had swollen.
Entonces se dio cuenta de que la madera seca de las ruedas se había hinchado.

But in fact we are not really interested in this pair of languages. Instead, we are interested in the pair Sense Tagged English - Spanish:

00117620r he 02154508v that the 02551380a 15098161n of the 04574999n had 00256507v . Entonces se dio cuenta de que la madera seca de las ruedas se había hinchado.

4.5 Word alignment algorithms

So, having a parallel corpus Sense Tagged English - Target Language, the task of deriving the local WordNet can be viewed as a word alignment problem. We need an algorithm capable to select the following relations (from the example in 4.4)

Synset	- Spanish	- Catalan
00117620r	- entonces	- llavors
02154508v	- darse cuenta	- adonar-se
02551380a	- seco	- sec
15098161n	- madera	- fusta

Fortunately, word alignment is a well-known task and there are several algorithms available to solve it. In this work we will test two algorithms:

- MFT: A very simple algorithm based on the detection of the most used target word of the same part-of-speech in the translation of the sentences containing the PWN synset.
- Berkeley Aligner⁴ (Liang et al., 2006): A well known word alignment algorithm.

⁴<http://code.google.com/p/berkeleyaligner/>

In both algorithms we force two restrictions: (i) we only detect as a variant for a given synset simple lexical units, that is, no multiwords; and (ii) we only detect one variant for each synset. In a future work we will try to overcome such restrictions.

4.5.1 MTF - Most Frequent Translation

This algorithm works as follows:

- We take all the synsets in the corpus ordered by frequency, beginning with the most frequent.
- For every synset we take all target sentences aligned to the corresponding source sentences containing this synset.
- The most frequent target lemma of the same POS from this set of sentences is the candidate variant for the synset.

For example, the synset 00393105a (*white_I*) appears 1190 times in the corpus. By applying the algorithm we get a set of candidates. If we take the most frequent adjectives in the translation of the sentences containing this synset, we obtain for Spanish:

blanco:1139; pequeño:212; perenne:160;
grande:105; rosa:86; rojo:83; amarillo:81;
fragante:66; púrpura:60; azul:53;

The most frequent is *blanco*, that is several times more frequent than the second one. So we take this word as a variant in Spanish for this synset.

4.5.2 Berkeley Aligner

In our experiments we also used the Berkeley Aligner (Liang et al., 2006). We used it with default options and only one variant for each synset was taken -the one bearing the higher probability. In these experiments we used lemma and part-of-speech information.

4.6 Linguistic resources

For our experiments we need a sense tagged corpus. The tags must be the PWN 3.0 synsets. Fortunately there are two freely available resources:

- The Semcor corpus⁵ (Miller et al., 1993).
- The Princeton WordNet Gloss Corpus (PWGC)⁶, consisting of the WordNet 3.0 glosses with semantic annotation.

⁵<http://www.cse.unt.edu/~rada/downloads.html>

⁶<http://wordnet.princeton.edu/glosstag.shtml>

Corpus	Sentences	Words
Semcor	37.176	721.622
PWGC	117.659	1.174.666
Total	154.835	1.896.288

Table 1: Size of the corpus

Spanish		Catalan	
Synsets	Variants	Synsets	Variants
T	115.989	69.829	70.856
N	80.324	48.676	51.598
V	16.805	9.010	11.577
A	17.752	11.450	7.679
R	1.108	693	2

Table 2: Size of the preliminary WordNets used for evaluation

In table 1 we can observe the total number of sentences and words in the corpus.

4.7 Linguistic tools

4.7.1 Machine Translation Systems

- Google Translate⁷.
- Microsoft Bing Translator⁸.

Both systems are statistical and allow us to translate from English to Spanish and Catalan.

4.7.2 Morphosyntactic tagger

All sentences in the corpus, both source and target sentences, have been morphosyntactically tagged with Freeling (Padró et al., 2010a).

4.7.3 Preliminary versions of the Spanish and Catalan 3.0 WordNets

The results obtained with the different algorithms have been automatically evaluated with a preliminary version of the Spanish and Catalan 3.0 WordNets. These preliminary versions have been obtained from versions 1.6 applying a mapping and other techniques. In table 2 we can see the number of synsets and variants of these versions.

5 Evaluation

We conducted the experiments for three language pairs:

- English - English: this dummy language pair allows to evaluate the effect of the quality of the machine translation system in the results. As we are using the source language we simulate a perfect machine translation system.

⁷<http://translate.google.com>

⁸<http://www.microsofttranslator.com/>

- English - Spanish

- English - Catalan

We have tested all the algorithms for every language pair. As machine translation systems, for English-Catalan and English-Spanish we have used both Google Translate and Microsoft Translator.

5.1 Algorithm MFT

In figure 1 we can see the results of this algorithm for all the language pairs (for the translated pairs we show only the results for Google Translate). In the figure we can see the precision obtained regarding the number of translated synsets. As we can realize, precision decreases along with the number of synsets. As we start the evaluation with the most frequent synset, this means that the precision decreases with the decrease of frequency, as it was expected.

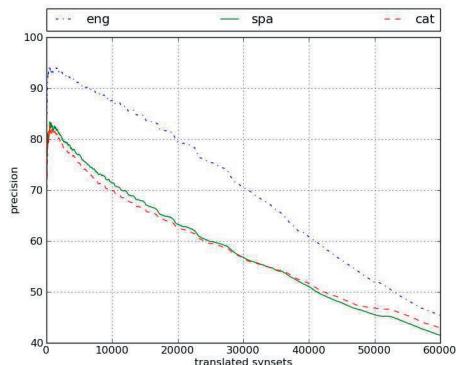


Figure 1: Precision of MFT for English, Spanish and Catalan (using Google Translate).

As we can see in the figure, with this algorithm we can obtain for the pair English-English about 20.000 translated synsets with a precision of 80%, and about 30.000 with a precision score of 70%. For this language pair we don't have the negative effects of errors coming from the machine translation system. As we can observe, results for Spanish and Catalan are very similar so that we can obtain about 2.200 translated synset with 80% precision and 10.000 with 70%.

We must keep in mind that we are performing an automatic evaluation using the full PWN 3.0 for English but using preliminary versions for Spanish and Catalan. Therefore, in some cases we

might be rejecting correct variants, as they are not present in the preliminary versions. For English this is less likely to happen as PWN is much more complete.

Results obtained with Microsoft Translator are very similar with a very slight decrease in precision. Figure 2 shows a comparison for Spanish.

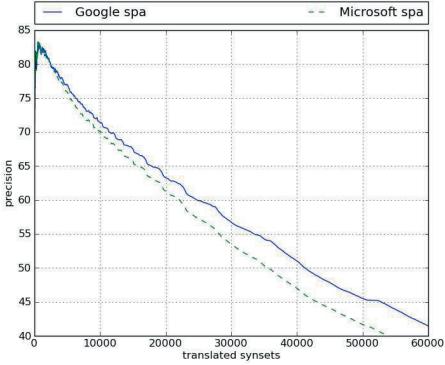


Figure 2: Comparison of MFT for Spanish using Google Translate and Microsoft Translator

5.1.1 Berkeley Aligner

Figure 3 shows the results achieved by using the Berkeley Aligner -please notice the change of x and y scales respect figure 1. In the figure we can observe some interesting facts. First of all, precision has increased significantly for the pair English-English so that it remains practically constant regardless of the frequency of the synset. For this ideal language pair we obtain precision scores higher than 96%. For real language pairs results are not so good since precision decreases with frequency. We can also observe that results for Spanish are better than for Catalan and the difference is higher for smaller frequencies.

As we are using a statistical word alignment algorithm, we can get a probability score for each synset. We can improve precision by taking only those alignments with a probability higher than a given threshold. Of course, this will have a cost on the recall but the goal is to get a subset of results indicating high precision. We have taken 0.9 as probability threshold, that is, we get the alignment if its probability is 0.9 or higher. In figure 4 we can observe the results for Spanish and a comparison with MFT and Berkeley Aligner taking the best candidate for all alignments, regardless of its precision. As we can observe, Berlkeley Aligner per-

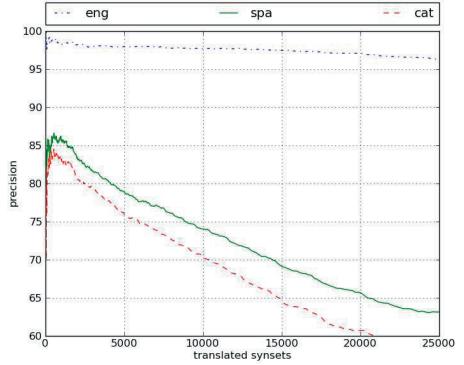


Figure 3: Precision for Berkeley Aligner for English, Spanish and Catalan.

forms better than MFT for all frequencies -about 5% better. Taking those alignments with a precision of 0.9 or higher we get a good improvement of the results until the 15.000th synset. Until the 5.000th synset the improvement is about 7 points with respect the Berkeley Aligner.

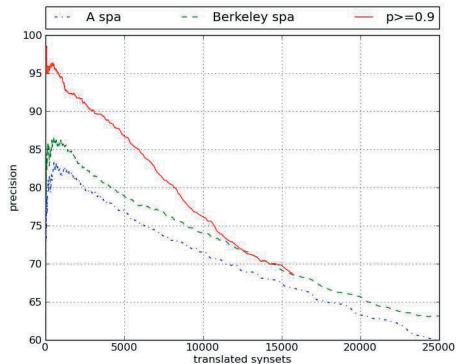


Figure 4: Comparison of MFT, Berkeley Aligner and Berkeley Aligner for $p \geq 0.9$ for Spanish

In figure 5 we can see the same comparison for Catalan. In this case Berkeley Aligner starts with better results but for a given frequency MFT starts to perform better. If we see the results for the Berkeley Aligner with probabilities higher than 0.9, we can observe that we get better results until the 10.000th synset. Until the 5.000th synset the improvement is also about 7 points with respect the Berkeley Aligner.

As for MFT results for Berkeley Aligner for the different machine translation systems are practically identical. In figure 6 we can see a compar-

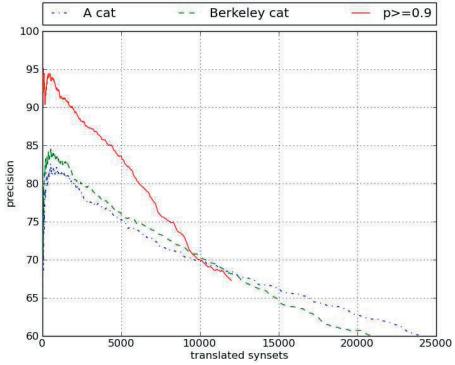


Figure 5: Comparison of MFT, Berkeley Aligner and Berkeley Aligner for $p \geq 0.9$ for Catalan

ison for Spanish using Google Translate and Microsoft translator. As we can see we obtain slight better results with Google Translate. The same applies for Catalan.

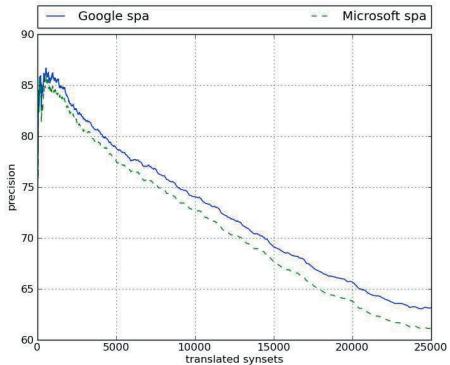


Figure 6: Comparison of Berkeley Aligner for Spanish using Google Translate and Microsoft Translator.

6 Conclusions and future work

In this paper we have presented a methodology for the construction of WordNets based on machine translation of a sense disambiguated corpus of English. This methodology has been used for the construction of the Spanish and Catalan WordNets 3.0 along with methodologies based on a mapping of previous existing versions.

This methodology achieves values of precision and number of obtained synsets comparable to methodologies based on bilingual dictionaries

for Spanish (Atserias et al., 1997) and Catalan (Benítez et al., 1998). To perform a good comparison we need to further analyse our results in order to group synsets according to the degree of polysemy. For Spanish, our best algorithm (Berkeley Aligner for $p \geq 0.9$) performs better than all the criteria presented in (Atserias et al., 1997) except monosemic-1 criterion. Nevertheless, our proposal performs worse than their combination of criteria as they can obtain 7.131 with a precision higher than 85% whereas we can obtain only 5.890 in the same conditions. For Catalan (Benítez et al., 1998) obtained much better results and our best proposal only outperforms monosemic-3 and all polysemic criteria.

This methodology can be used for the construction of new WordNets and also for the enrichment of existing WordNets. Our proposal can be applied to those languages having an available English-source machine translation system. Such a system must be able to perform good lexical selection while other translation errors are less important. It has to be noticed that Google Translate offers machine translation from English to more than 50 languages and Microsoft Translator to more than 35 languages. Of course, not all language pairs achieve the same quality, but they can be a good starting point.

In future work we plan to overcome some of the restrictions of the algorithms presented in this paper. First of all, we will use Berkeley Aligner to try to get more of one variant for each synset. This can be done by observing the assigned probability and taking more than one candidate if probability scores are similar enough. We will aim to overcome as well the restriction of considering the possible variants only when they are simple lexical units. This can also be done by further exploring the possibilities of the alignment algorithm.

Finally, we want to cope with the limit of recall due the corpus: only those synsets present in the corpus can be retrieved. To achieve this goal we will explore several possibilities, including the use of automatic semantic tagging (Padró et al., 2010b) and the use of Wikipedia page links as a pseudo-semantical tagging as in (Navigli and Ponzetto, 2010).

References

- E. Agirre, O. Ansa, X. Arregi, J. M. Arriola, A. D. de Iarrazá, E. Pociello, and L. Uriá. 2002. Method-

- ological issues in the building of the basque WordNet: quantitative and qualitative analysis. In *Proceedings of the first International WordNet Conference in Mysore, India*, page 21–25.
- J. Atserias, S. Climent, X. Farreres, G. Rigau, and H. Rodriguez. 1997. Combining multiple methods for the automatic construction of multi-lingual WordNets. In *Recent Advances in Natural Language Processing II. Selected papers from RANLP*, volume 97, page 327–338.
- I. Azarova, O. Mitrofanova, A. Sinopalnikova, M. Yavorskaya, and I. Oparin. 2002. Russnet: Building a lexical database for the Russian language. In *Workshop on WordNet Structures and Standardisation, and how these affect WordNet Application and Evaluation*, pages 60–64, Las Palmas de Gran Canaria (Spain).
- Laura Benítez, Sergi Cervell, Gerard Escudero, Mònica López, German Rigau, and Mariona Taulé. 1998. Methods and tools for building the catalan WordNet. In *In Proceedings of the ELRA Workshop on Language Resources for European Minority Languages*.
- R. L Cilibrai and P. M.B Vitanyi. 2007. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383.
- I. Dagan and A. Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.
- C. Fellbaum. 1998. *WordNet: An electronic lexical database*. The MIT press.
- Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL '06*, page 104–111, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1220849.
- Palmira Marrafa. 2002. Portuguese WordNet: general architecture and internal semantic relations. *DELTa: Documentação de Estudos em Lingüística Teórica e Aplicada*, 18(spe).
- M. Miháltz, C. Hatvani, J. Kuti, G. Szarvas, J. Csirik, G. Prószyk, and T. Váradi. 2008. Methods and results of the Hungarian wordnet project. In *Proceedings of the Fourth Global WordNet Conference. GWC*, pages 387–405, Szeged, Hungary.
- George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology, HLT '93*, page 303–308, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1075742.
- Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, page 216–225, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1858704.
- F. J Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.
- L. Padró, M. Collado, S. Reese, M. Lloberes, and I. Castellón. 2010a. FreeLing 2.1: Five years of open-source language processing tools. In *LREC*, volume 10, page 931–936.
- L. Padró, S. Reese, E. Agirre, and A. Soroa. 2010b. Semantic services in freeling 2.1: Wordnet and UKB. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*.
- E. Pianta, L. Bentivogli, and C. Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *1st International WordNet Conference*, pages 293–302, Mysore (India).
- D. Putra, A. Arfan, and R. Manurung. 2008. Building an Indonesian WordNet. In *Proceedings of the 2nd International MALINDO Workshop*.
- B. Sagot and D. Fišer. 2008. Building a free French wordnet from multilingual resources. In *Proceedings of OntoLex*.
- M. Savaški and I. Trajkovski. 2010. Automatic construction of wordnets by using machine translation and language modeling. In *13th Multiconference Information Society*, Ljubljana, Slovenia.
- D. Tufis, D. Cristea, and S. Stamou. 2004. BalkaNet: aims, methods, results and perspectives: a general overview. *Science and Technology*, 7(1-2):9–43.
- P. Vossen. 1996. Right or wrong: combining lexical resources in the EuroWordNet project. In *Proceedings of Euralex-96*, page 715–728, Goetheborg.
- P. Vossen. 1998. Introduction to Eurowordnet. *Computers and the Humanities*, 32(2):73–89.
- P. Vossen. 1999. EuroWordNet a multilingual database with lexical semantic networks. *Computational Linguistics*, 25(4).

Kannada Verbs and their Automatic Sense Disambiguation

S. Parameswarappa

Department of CS & E

Malnad College of Engg., Hassan.

param.phd@gmail.com

V.N. Narayana

Department of CS & E

Malnad College of Engg., Hassan.

vnnarayana@yahoo.com

Abstract

The present paper explores the use of argument structure for Kannada Verb Sense disambiguation. We argue that the argument structure of a verb will have a vital role in verb sense disambiguation. The context is the only means to identify the meaning of a polysemous word in a sentence. Hence, in the present work the arguments and their relationship with the verb in a sentence are considered as a context. The Shallow parser built by International Institute of Information Technology (IIIT) Hyderabad has been used for extracting the syntactic and morphological information of a sentence. We constructed a moderate size electronic dictionary for Kannada language. We developed a web corpus for testing the proposed Verb Sense Disambiguation algorithm. Thirty most frequently used polysemous verbs are carefully selected for disambiguation task. There are about eighty sentences extracted from a web corpus to test the algorithm. The proposed algorithm is implemented in Program Extraction and Reporting Language (Perl). The experiments are conducted and the results obtained are described. The efficiency of the algorithm is proved to be reliable and extendable. The proposed solution is useful for Kannada language processing task in general and Machine Translation in particular.

Key words: Argument Structure, Verbalizer, Dictionary, Verb Sense Disambiguation, Thematic role, Semantic pattern, Corpora.

1 Introduction

A word can have different meaning depending on the context in which it is used. Word Sense Disambiguation (WSD) is the task of determining the correct meaning (“sense”) of a word in a given context.

Let us consider a Kannada verb ಇಡು [iDu], it has three different meanings (senses). The senses and their examples for a verb ಇಡು [iDu] are depicted in Table 1.

Sense Label	Sense Definition	Example Sentence
iDu1	ರೇವಣೆ ಇಡು [Thee-vaNe iDu] ‘To deposit’	ರಾಮನ ಖಾತೆಗೆ ಬ್ಯಾಂಕಿನಲ್ಲಿ ಹಣ ಇಡು. [raamana khaatige byaankinalli haNa iDu] ‘Deposit money to rama’s account in a bank’
iDu2	ಅನ್ನ ಇಡು [anna iDu] ‘Serve food’	ರಾಮನಿಗೆ ಅನ್ನ ಇಡು. [raamanige anna iDu] ‘Serve rice to rama’
iDu3	ಹೆಚ್ಚಿ ಇಡು [hejjie iDu] ‘Set aside the foot’	ರಾಮನ ಜೊತೆ ಹೆಚ್ಚಿ ಇಡು. [raamana jote hejjie iDu] ‘go along with rama’

Table 1: The verb ಇಡು[iDu] and its meaning.

The present paper proposes the solution to disambiguate Kannada polysemous verbs like ಇಡು [iDu] using argument structure.

The study conducted on a Kannada dictionary with 46559 entries reveals the fact that the verbs seem to exhibit higher ratio of semantic ambiguity than other lexical categories. We found 314 ambiguous verbs out off 2202 verbs in a dictionary. The higher score of polysemy with verbs is an indication of how important verbs are in developing natural language applications. Frequently used verbs in Kannada such as ತಿನ್ನಿ [tin-nu] ‘to eat’, ಹೋಗು [hoogu] ‘to go’ and ಮಾಡು [maaDu] ‘to do’ are also most polysemous. Some of them function as verbalizers when used with nouns. In order to achieve higher quality translation output in machine translation, verb sense disambiguation is one of the most important

problems to be solved. This is the motivation behind the present work.

Each verb has finite number of distinct senses. The dictionary will list all of them. The verb predicate argument structure specifies the possible syntactic structure of a sentence in which it occurs. The linking of argument with thematic roles such as Agent, Patient, Theme, Experiencer, Benefactive, Goal, Source, Location etc determine different meaning or senses of the event or action described by the predicate. This kind syntactic and semantic information is generally considered as verb lexical property.

1.1 Organization of the paper

Section 2 explores the related work. Section 3 describes the Kannada language in general and Kannada verb and its argument structure in particular. Section 4 specifies the basic infrastructure requirements of the present work. Section 5 describes the methodology and the proposed algorithm used to disambiguate the Kannada verbs. Section 6 provides the detailed information required to implement the proposed algorithm. Section 7 describes the evaluation of the algorithm. Section 8 concludes the paper.

2 Related work

Baker (1988) formulates the Uniformity of Theta Assignment Hypothesis. According to Grimshaw (1990) arguments structure can be constructed in accordance with thematic hierarchy. Adele (1995) proposed that grammatical construction plays a central role in the relation between form and meaning in simple sentence. Taegoo (2000) in his work on argument structure and English grammar introduced the basic concepts about the arguments and argument structure, argument and thematic roles and argument and case.

Gruber (2001) works on thematic roles and grammatical arguments in a sentence are commonly described in terms of their relations. Paola and Suzanne (2001) conducted supervised learning experiments for automatic verb classification based on statistical distribution of argument structure. Hoa and Martha (2005) described an automatic word sense disambiguation system that disambiguate verb senses using syntactic and semantic features that encode information about predicate arguments and semantic classes. Kumar et al. (2005) conducted a comparative study of English and Hindi verb semantics. They ex-

plore syntactic and semantic variations while mapping English verbs onto Hindi verbs.

Jinying et al. (2006) conducted an empirical study of the behavior of active learning for word sense disambiguation. They show that two uncertainty based active learning methods combined with a maximum entropy model work well on learning English verb senses. Chen et al. (2009) presents a high performance broad coverage supervised WSD system for English verbs that uses linguistically motivated features and a smoothed maximum entropy machine learning model. Rafiya et al. (2008) developed verb frames for Hindi. They capture syntactic commonalities of semantically related verbs. Rafiya and Dipti (2010) introduced a preliminary work on Hindi causative verbs. They presented the morphology, semantic and syntax of the Hindi causative verbs. Matthew and Nick (2010) proposed a method to develop an inventory of verb argument construction based upon English form, function and usage.

3 The Kannada Language

Indian languages belong to four different language families namely, The Indo-Aryan, The Tibeto-Burman, The Austro-Asiatic and the Dravidian. In a present paper, we propose a solution for Kannada verb sense disambiguation using argument structure. Kannada belongs to Dravidian family (Kempegowda, 2008).

Even though, Kannada has very old and rich literary tradition, it is currently spoken by very large number of people (around 60 million), and Karnataka is in the centre stage of IT (Information Technology) revolution in a country, When it comes to technology, Kannada is one of the technologically least developed language in India today. As of today, the only corpus we have is the roughly 3 Million word corpus developed by Central Institute of Indian Languages (CIIL) Mysore long ago. Lack of basic resources such as corpora is one of the major reasons for our lagging behind in language technology (Murthy, 2001). To address the issue, we designed an algorithm and implemented it using Perl under Linux environment to construct web corpora automatically. We used selected sentences from web corpora for testing our verb sense disambiguation algorithm.

Kannada is an agglutinating language of suffixing type. Nouns are marked for number and case and verbs are marked, in most cases, for agreement with the subject in number, gender

and person. Therefore Kannada is a relatively free word order language (Sridar, 2007).

Kannada language exhibits a very rich system of morphology. Morphology includes inflection, derivation, conflation (sandhi) and compounding.

3.1 Argument Structure

The concept of argument structure is borrowed from logic. It is generally concern with relations between predicate and a set of arguments (Murthy and Keshva, 2002). The crucial element of a sentence in Kannada is predicate, which is usually a verb or noun. The present study will be limited to verbal predicates only. The predicate determine the presence or absence of other crucial elements in a sentence.

Consider a sentence ರಾಮ ನಿನ್ನ ಹಣ್ಣನ್ನ ತಿದನು. [raama ninne haNNanu tindanu] ‘rama ate fruit yesterday’ the constituents or elements of the above sentence are ರಾಮ [raama] (Subject), ನಿನ್ನ [ninne] (Adverb), ಹಣ್ಣ [haNNu] (Object), ತಿನ್ನ [tinnu] (Verb). Among these some elements are obligatory and others are optional. Table 2 shows the possible derivations for the example sentence.

S N	Sentence Derivations	Comment
1	ನಿನ್ನ ಹಣ್ಣನ್ನ ತಿದನು. [ninne haNNanu tindanu] ‘Yesterday (he) ate fruit’	explicit subject is missing
2	ರಾಮ ನಿನ್ನ ಹಣ್ಣನ್ನ [raama ninne haNNanu] ‘rama yesterday fruit’	verb is missing
3	ರಾಮ ನಿನ್ನ ತಿದನು. [raama ninne tindanu] ‘rama ate yesterday’	object missing
4	ರಾಮ ಹಣ್ಣನ್ನ ತಿದನು. [raama haNNanu tindanu] ‘rama ate fruit’	adverb is missing

Table 2: Derivations for example sentence.

Among the above listed sentence derivations only the sentence two is ungrammatical and all others are grammatical. With this illustration, we can show that only the verb in the example sentence is obligatory and other elements are optional. The obligatory elements holds maximum amount of information in a sentence. The elements which are required by the predicate may be called arguments. Every predicate has its own set of arguments defined by its semantic properties. The syntactic structure of the sentence or the

clause of which the predicate is the head is determined by the semantic property and its argument structure. Based on the number of arguments required by the predicate, the predicates are classified in to three types, namely one-place, two-place, three-place predicates, they requires one, two and three arguments respectively(Gowda, 2008). Table 3 shows the examples for them.

Predicates	Examples
One place	ರೈಲು ಹೋಗುತ್ತಿದೆ. [rylu hooguttide] ‘Train is moving’
Two place	ರಾಮನು ಹಣ್ಣನ್ನ ತಿದನು. [raamanu haNNanu tindanu] ‘rama ate fruit’
Three place	ರಾಮನು ಸಿಂಗೆ ಜಿಂಕೆಯನ್ನ ಕೊಟ್ಟನು. [raamanu siitege jinkeyannu koT-Tanu] ‘rama has given deer to siite’

Table 3: Types of predicates.

3.2 Arguments and Thematic Roles

The predicate occupy a salient position in the sentence. It defines relationships such as who is doing the action and who or what is being affected by the action.

Consider an example ರಾಮ ಹಣ್ಣ ತಿದ. [raama haNNu tinda] ‘rama ate the fruit’. In the above sentence, ರಾಮ [raama] ‘rama’ acts as agent of the action denoted by the verb and the ಹಣ್ಣ [haNNu] ‘fruit’ acts as affected object. Such relations are generally known in the literature as thematic relations (Gruber, 2001). These are also called as semantic relations. Within the principles and parameters of language theory, these relations are treated under Theta theory. Here, every lexical entry for a verb must specify a set of theta role that occur with it. The relationship between the theta role and predicate is captured by the logical expression in the following way.

ತಿನ್ನ (ರಾಮ,ಹಣ್ಣ) [tinnu (raama,haNNu)] ‘ate(rama,fruit)’ which indicates that two arguments ರಾಮ [raama] ‘rama’ and ಹಣ್ಣ [haNNu] ‘fruit’ are related by the predicates semantic property.

The well known distinction of verbs into transitive and intransitive is based on argument structure of the predicates. If the verb takes one argument, then it is called as intransitive verbs and there is no provision for object. Otherwise, if it takes two or more arguments, then it forms a

transitive construction, where the subject and object are provided (Bagum et al., 2008).

Semantic patterns of arguments are captured through thematic roles (Rappaport and Levin, 1988). The following thematic roles assigned to each argument of the predicate.

- a) Agent: An entity, which intentionally instigates the event or an action described by the predicate. Example: ಅಪ್ಪೆ ನನಗೆ ಹಣ ಕೊಟ್ಟರು [appa nanage haNa koTTaru] ‘father has given me money’ (appa is an agent).
- b) Patient: An entity, which undergoes the effect of event or affected by the event or an action described by the predicate. Example: ರಾಮು ಹುವು ಕಿಟ್ಟಿದ್ದಾನೆ [raama huuvu kittiddaane] ‘Rama peeled the flower’ (flower is a patient).
- c) Theme: An entity, which is moved in the event or an action described by the predicate. Example: ರಾಮು ಮನೆ ಕಟ್ಟಿ [raama mane kaTTida] ‘rama built a house’ (mane is a theme).
- d) Experiencer: An entity, which experiences some psychosomatic state described by the predicate. Example: ರಾಮನಿಗೆ ಜ್ವರ ಇದೆ [raamanige jvara ide] ‘Rama has fever’ (raama is an experiencer).
- e) Instrument: An entity, which is used to realize an action or an event or an action described by the predicate. Example: ನಾನು ಕಾಕುವಿನಿಂದ ಕಡ್ಡಿಸಿದೆ [naanu caakuvininda haNNanu kattariside] ‘I cut the fruit with a knife’ (caaku is an instrument).
- f) Locative: The place in which an event or an action described by the predicate takes place. Example: ರಾಮು ಕಾಡಿನಲ್ಲಿ ವಾಸವಾಗಿದ್ದು [raama kaadinalli vaasavaagidda] ‘Rama lived in forest’ (kaadu is a locative).
- g) Goal: Entity towards which something moves in an event or an action described by the predicate. In some classification goal is distinguished from beneficiary or recipient. Example: ಅಪ್ಪೆ ನನಗೆ ಹಣ ಕೊಟ್ಟರು [amma nanage haNa koTTaru] ‘mother has given me money’ (nanage is a goal).
- h) Source: Entity, from which something moves in an event or an action described by the predicate. Example: ರಾಜು ಬೆಂಗಳೂರಿನಿಂದ ಬಂದು [raaju banenorininda banda] ‘Raju came from Bangalore’ (Bangalore is a source).

3.3 Types of Arguments

Based on the syntax, the arguments are classified into following types (Williams, 1981).

- a) External and Internal arguments: The argument, which is associated with the position outside the maximal projection of the predicate, is called as external argument where as that of

within the maximal projection is called as internal arguments. Example: ರಾಮು ಉಟ ಮಾಡಿದ [raama uuTa maaDida] ‘Rama ate food’. Here, Rama is an external argument and uuTa is an internal argument.

- b) Direct and Indirect argument: Except external arguments, if an argument is realized with a post position then it is called as indirect argument otherwise, it is a direct argument. Example: ರಾಮನು ಶಿತೇಗೆ ಜಿಂಕೆಯನ್ನು ಕೊಟ್ಟಿದು [raama siitege jinkeyannu koTTanu] ‘Rama gave deer to siite’. Here siite is an indirect argument and jinke is a direct argument.

3.4 Argument Structure & Kannada verbs

Intransitive verbs do not form a homogeneous group in Kannada. We can have distinct sub types based on their semantic features (Sridar, 2007).

- a) Unergative verbs: They form a special group of intransitive verbs. Semantically, unergative verbs have subject perceived as actively initiating or actively responsible for an action expressed by the verb. Thematically, these verbs take an agent, which is an external argument. The following are some of unergative verbs identified in Kannada. ಕಾಗು [kuugu] ‘to shout’, ತಾಡು [iiju] ‘to swim’, ಅಳು [aLu] ‘to cry’.
- b) Unaccusative verbs: There exists a group of intransitive verbs, characterized semantically, where the subject does not actively initiate or is not actively responsible for an action of the verb, rather it has proportion which it shares with the direct object of a transitive verb. Thematically, these verbs take Theme, which is an internal argument in terms of argument structure. The following are some of the unaccusative verbs identified in Kannada. ಕುದಿಸು [kudisu] ‘to boil’, ತಲುಪು [talupu] ‘to reach’.

4 Experimental setup

The argument structure is a base for our experimentation. The argument structure is used to disambiguate the Kannada verb. The details of a corpus, dictionary and parser used for experimentation are given below.

4.1 Corpora

A corpus is a large and representative collection of language material stored in a computer processable form (Sinclair, 1991). Open, freely and publicly available corpora can be used by all researchers as standard data sets to develop and test their system. A large and representative cor-

pus has many uses and applications (Kumar, 2007). What all we can do with a corpus is limited by our imagination and creativity. A corpus forms the very basic of all language and linguistic studies.

In the present work we have used randomly selected set of sentences from the Kannada web corpora developed by us for experimentation. We automated the complete corpora building process. An algorithm is designed and it is implemented in Perl to make corpora building process automatic. The corpora includes a wide variety of subject, such as recent discussions, blogs, articles, recent activities, proverbs, recent feedbacks, poem, fifteen books, novels, news paper and informal chats (Parameswarappa et al., 2012).

4.2 Dictionary

Knowledge of language is essential for meaningful communication through language. Words of a language and the phonological, morphological, syntactic and semantic information associated with them forms a very important part of knowledge of a language. Knowing the words is an extremely important part of knowing a language. Dictionaries are storehouse of such information and therefore they have key role to play in Natural Language Processing (Murthy, 1997).

A Kannada electronic dictionary containing 46559 entries have been developed for our work. Each entry is on separate line. Each entry starts with the head word followed by tags separated by double vertical lines. Additional information fields, if any come at the end of each tag separated by double colons. Comments come at the end after hash. The dictionary is a single plain text file, amenable for manipulation through basic commands and tools such as grep, awk and sed. It is easy to write Perl scripts too. Internally, the dictionary will reside in an indexed m-way balanced tree structure.

4.3 Parser

The syntactic and morphological information of a sentence are extracted using the Kannada Shallow Parser available at IIIT, Hyderabad. (<http://ltrc.iiit.ac.in/analyzer/kannada/>).

5 Experimentation

This section describes the methodology and proposed algorithm for verb sense disambiguation.

5.1 Methodology

We describe the use of argument structure to disambiguate Kannada polysemous verbs. The sense of a verb is disambiguated by extracting the verb and its arguments with their semantic features from a given sentence. A match with the relevant cluster of arguments and an argument structure frame of a verb results in the identification of a correct sense.

5.2 Algorithm

Algorithm VERB-SENSE-DISAMBUGUATOR. Given an input sentence S, verb argument structure frame file V_Arg_Str_frm_FILE and verb arguments semantic features file V_Arg_Sem_Fea_FILE. M_FILE contains morphological information of all words in a sentence S. V is a polysemous verb to be disambiguated. VA_LIST contains list of arguments pertaining to V. VA is an individual arguments in VA_LIST. V_MEAN is an output variable provides exact meaning of V. This algorithm disambiguates the polysemous verb V in an input sentence S using argument structure.

1. [Input a sentence]
Read a sentence S.
2. [Shallow Parsing]
Perform shallow parsing on a sentence S to extract morphological information of all words in S and store it in M_FILE.
3. [Extract verb and its arguments from M_FILE]
Extract verb from M_FILE store it in V.
Extract all verb arguments from M_FILE store them in VA_LIST.
4. [Sense Disambiguation]
If V found in V_Arg_Str_Frm_FILE then
Begin
for each VA in VA_LIST do
Begin
Separate verb sense meaning from features store it in V_MEAN.
Extract features based on verb sense and concatenate them with VA.
if VA found in V_Arg_Sem_Fea_FILE then
Next VA
End
End
5. [output correct verb sense]
print(V_MEAN)

6 Implementation

The algorithm is implemented using Perl. The required files, list and the modules for implementing the algorithm are discussed in this section.

6.1 Files and List

a) V_Arg_Str_frm_FILE: Verb Argument Structure Frame file is an input file. It contains possible different senses of a verb with its arguments frame structure. Fig. 1 shows the partial content of a file for a verb ಒಣಗು [oNagu]

oNagu, v, [to dry, to heal]
oNagu, {[N<Th>(-a, +c)]}; "dry"
oNagu, {[N<Ex>(+)#[N <Th>Q]}; "heal"

Figure 1. V_Arg_Str_frm_FILE content.

b) V_Arg_Sem_Fea_FILE: Verb Arguments Semantic Feature file is an input file. It contains semantic features of verb arguments. Fig. 2 shows the partial content of a file for the following example sentences.

ಒಟ್ಟೆಗೆಲು ಒಣಗಿವೆ.
[baTTegaLu oNagive] 'cloths have dried'.
ಅವನ ಗಾಯ ಒಣಗಿದ್ದ.
[avana gaaya oNagide] 'his wound has healed'.

baTTegaLu, N(-a,+c)
avana, P(+h)
gaaya, N()

Figure 2. V_Arg_Sem_Fea_FILE content.

c) M_FILE: it is a temporary file created after shallow parsing the input sentence S. It provides morphological information of all words. Fig. 3 shows the content of a file for the following example sentence.

ಒಟ್ಟೆಗೆಲು ಒಣಗಿವೆ.
[baTTegaLu oNagive] 'cloths have dried'.

baTTegaLu NN <fs af='baTTe,n,n,pl'>
oNagive VM <fs af='oNagu,v,n,pl'>

Figure 3. M_FILE content.

d) VA_LIST: It contains all the arguments pertaining to verb V. The contents of VA_LIST for a verb ಒಣಗು [oNagu] and the above cited example sentence are ಒಟ್ಟೆಗೆಲು [baTTegaLu] 'cloths', ಅವನ [avana] 'his' and ಗಾಯ [gaaya] 'wound'.

Fig. 1-3 shows only the sample example entries. In the similar way, we encoded thirty potential ambiguous Kannada verbs and stored

them in V_Arg_Str_frm_FILE. The program is tested on eighty Kannada sentences. The arguments semantic features pertaining to verb in a sentence is stored in V_Arg_Sem_Fea_FILE.

6.2 Implementation modules

a) Verb and Arguments Extractor: This module extract polysemous verb V and its arguments from a sentence S.

b) Sense Disambiguator: This module will identify and extract correct sense of a polysemous verb V.

7 Evaluation

We use the sentences extracted from Kannada web corpora for testing the program.

7.1 Test document

Table 4 shows the partial list of sentences used to test the program.

ಒಟ್ಟೆಗೆಲು ಒಣಗಿವೆ. [baTTegaLu oNagive] 'cloths have dried'.
ಅವನ ಗಾಯ ಒಣಗಿದ್ದ. [avana gaaya oNagide] 'his wound has healed'.
ಅವನ ಮದುವೆ ಜರುಗಿದ್ದ. [avana maduve jarugide] 'his marriage took place' .
ಆಕೆ ಪಕ್ಕಾಗಿ ಜರುಗಿದ್ದಳು. [aake pakkakke jarugidaLu] 'she moved aside'.
ಮುಕ್ಕಳು ಕಾಗಡಗಳನ್ನು ಸುಳಿತು. [makkaLu kaagadagaLanu suTTaru] 'children burnt the papers'.
ಅವನು ರೊಟ್ಟಿ ಸುಳಿತು [avanu roTTi suTTanu] 'he roasted the bread'.
ರಾಮನ ಖಾತೆಗೆ ಬ್ಯಾಂಕಿನಲ್ಲಿ ಹಣ ಇಡು. [raamana khaatenge byaankinalli haNa iDu] 'Deposit money to rama's account in a bank'
ರಾಮನಿಗೆ ಅನ್ನ ಇಡು. [raamanige anna iDu] 'distribute rice to rama'
ರಾಮನ ಜೋತೆ ಹೆಚ್ಚಿ ಇಡು. [raamana jote hejje iDu] 'go along with rama'

Table 4: Test sentences.

7.2 Result

The results obtained for the test sentences are shown in Table 5. Even though, the program disambiguate all the sentences listed in Table 4 correctly out of 80 test sentences the program disambiguate only 60 sentences correctly. The

errors are due to morphological analysis, discourse level ambiguity etc.

Verb	Example sentence	Sense
ಒಣಗು	ಬಟ್ಟಿಗಳು ಒಣಿವೆ.	Dry
	ಅವನ ಗಾಯ ಒಣಿದೆ.	Heal
ಜರುಗು	ಅವನ ಮುದುವೆ ಜರುಗಿದೆ.	Happen
	ಆಕೆ ಪಕ್ಕೆ ಜರುಗಿದಳು.	Move
ಸುಡು	ಮುಕ್ಕಳು ಕಾಗದಗಳನ್ನು ಸುಡುತ್ತು.	To burn
	ಅವನು ರೊಟ್ಟಿ ಸುಡುತ್ತು	Roast
ಇಡು	ರಾಮನು ಖಾತೆಗೆ ಬ್ಯಾಂಕನಲ್ಲಿ ಇಡುತ್ತು.	To deposit
	ರಾಮನಿಗೆ ಅನ್ನ ಇಡು.	Serve food
	ರಾಮನು ಚೀಲೆ ಹೆಚ್ಚಿ ಇಡು.	Set aside the foot

Table 5: Program Results.

7.3 Observation

During the process of verb sense disambiguation, the following points are noticed.

- The creation of Verb Argument Structure Frame file and Verb Arguments Semantic Feature file will have a critical role in the disambiguation process. If these two files provides the exhaustive information then the performance of the proposed algorithm is guaranteed to be high.
- Manually creating above cited files are labor intensive and time consuming.
- If the files creation process is automated then the proposed solution is reliable.
- The verb sense disambiguation task highly depended on the lexical and syntactic information along with semantic information, hence a good parser will also plays major role during the disambiguation process.
- Knowledge of an argument structure and the thematic roles assigned by the Kannada verb to its arguments solely contribute to the understanding of sentences for verb sense disambiguation task.
- Attempt have been made to deal with the Kannada verbs in detail with respect to their syntax and semantics.
- The output generated by the proposed system for a sentence ಸೀತೆ ನೆಡಿರು [siite neredaru] is wrong. Because, the system assign a sense 'Seethe gathered' instead of 'Seethe matured bio-

logically'. The output generated by the system is wrong for this kind of sentences.

8 Conclusion and Future work

In this paper, we have explored the use of argument structure for Kannada verb sense disambiguation. The argument structure is the most significant component of the grammar that acts as an interface between syntax and semantics of the language. We investigated the argument structure of Kannada verbs and shown by implementation that argument structure can be used effectively to disambiguate verb senses. However, the present work does not address the problem of each and every verbs in Kannada that have different senses exhaustively. It is a frame work to build a robust verb sense disambiguation system for Kannada. To the best our knowledge and belief, it is a first attempt towards building automatic Kannada Verb Sense Disambiguation system.

The major limitation of present work is that the arguments in Kannada must be exhaustively analyzed and marked for their semantic features. It is certainly a difficult and time consuming task but will have greater gain in the long run particularly in the area of Natural Language Processing in general and Machine Translation in particular. Hence, in future, we are planning to build a robust Verb Sense Disambiguation System for Kannada by investigating the automatic argument semantic feature marking techniques.

References

- Adela E Goldberg. 1995. *A Constructive Grammar Approach to Argument Structure*. London, The University of Chicago press.
- Baker M. 1988. *A Theory of Grammatical function changing*. Chicago, The University of Chicago Press.
- Bharadwaja Kumar G, Kavi Narayana Murthy and Chaudhuri B.B. 2007. *Statistical Analysis of Telugu Text Corpora*. IJDL. Vol 36, No 2, pp – 71-99.
- Chen, Jinying, and Martha Palmer. 2009. *Improving English Verb Sense Disambiguation Performance with Linguistically Motivated Features and Clear Sense Distinction Boundaries*. Language Resources and Evaluation. 43:181–208, Springer Netherland: SemEval2007.
- Grimshaw J, A. Mester. 1990. *Argument Structure*. Cambridge Massachusetts, MIT Press.
- Gruber S Jeffery. 2001. *Thematic Relations in Syntax*. In the hand book of contemporary syntactic Theory

- (ed) Mark Baltin and Chric Collins, Massachusetts, Black well publication.
- Hoa Trang Dang, Martha Palmer. 2005. *The Role of Semantic Roles in Disambiguating Verb Senses*. Proceedings of the 43rd Annual Meeting of the ACL, pages 42–49, Ann Arbor, June 2005.
- Jinying Chen, Andrew Schein, Lyle Ungar, Martha Palmer. 2006. *An Empirical Study of the Behavior of Active Learning for Word Sense Disambiguation*. Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 120–127, New York, June 2006.
- Kavi Narayana Murthy. 1997. *Electronic Dictionaries and Computational Tools*. Linguistics Today, Vol 1, No 1, pp 34-50.
- Kempegowda K. 2008. *Thowlanika Draveeda Vyakarana (Comparative Dravidian Grammar)*. Gajana Publication, Mysore, India.
- Kittel F. 1903. *A grammar of the Kannada Language in English*. Basel mission book and track depositary. Mangalore. Digitized by Microsoft. University of Toronto Library.
- Kumar Sanjeev, Rafiya Begum, Radhika Mamidi. 2005. *A Comparative study of English and Hindi verb semantics*. LSI Platinum Jubilee Conference 6th-8th December, 2005 HCU.
- Kushaalappa Gowda K. 2008. *Kannada Bhashe & VyakaranagaLa Ondu samikshe (Kannada Language and its Grammar)*. Karthik enterprises. Bangalore. India.
- Matthew Brook O'Donnell, Nick Ellis. 2010. *Towards an Inventory of English Verb Argument Constructions*. Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics, pages 9–16, Los Angeles, California, June 2010.
- Murthy M.C. Kesava. 2002. *Argument Structure of Telugu Verbs*. ICOSAL-4. (mimeo) Annamalai University: Annamalainagar.
- Narayana Murthy K. 2001. *Computer Processing of Kannada Language*. University of Hyderabad, Hyderabad. In the workshop at Kannada University, Hampi.
- Paola Merlo, Suzanne Stevenson. 2001. *Automatic Verb Classification Based on Statistical Distributions of Argument Structure*. Computational Linguistics Volume 27, Number 3 2001, Association for Computational Linguistics
- Parameswarappa S and Narayana V.N 2012. *Novel Approach to Build Kannada Web Corpus*. Accepted oral presentation paper. Proceedings of ICCCI – 2012. Coimbatore. India.
- Rafiya Begum, Samar Husain, Lakshmi Bai and Dipti Misra Sharma. 2008. *Developing Verb Frames for Hindi*. Proc. of LREC08, 2008.
- Rafiya Begum, Dipti Misra Sharma. 2010. *A Preliminary Work on Causative Verbs in Hindi*. Eighth Workshop on Asian Language Resources (ALR8) held in conjunction with The 23rd Conference on Computational Linguistics (COLING 2010).
- Rappaport M & Levin B. (1988). *What to do with Theta - Roles*. In W.Wilkins (ed). Syntax and Semantics 21: Thematic Relations. New York: Academic Press.
- Sinclair J. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.
- Sridhar S.N. 2007. *Modern Kannada Grammar*. Manohar Publishers & Distributors. New Delhi. India.
- Taegoo Chung 2000. Argument structure and English Grammar. Korea University.
- Tenny C. 1992. *The Aspectual Interface Hypothesis*. Lexical matters, CSLI lecturer notes 24, Centre for the study of Language Information, Stanford, Stanford University.
- UmaMaheswara Rao.G. 2000. *Morphological Complexity in Telugu*. (mimeo). University of Hyderabad, Hyderabad.
- Williams E. 1981b. *Argument Structure and Morphology*. The Linguistic Review. 1, 81-114.

Wordnet and SUMO for Sentiment Analysis

Adam Pease, John Li, Karen Nomorosa

Rearden Commerce

Foster City, CA, USA

[adam.pease | john.li | karen.nomorosa]@reardencommerce.com

Abstract

In this paper, we show how Princeton's WordNet and associated resources can be used as part of an integrated system for sentiment analysis, called SigmaSentiment. We discuss the development of a system for sentiment analysis and concept extraction. We first provide an introduction to the user experience to motivate our work. We begin technical discussion with some background on the Suggested Upper Merged Ontology and WordNet, as well as their relation and the associated research that makes this work possible. We describe system components and data flow in detail, and some detail about the deployment architecture in the fielded system.

1 Introduction

Sentiment analysis refers to the assessment of the emotional content of text. Our goal is to support a more personalized travel experience in searching for a hotel. In this pilot project we combined techniques in computational linguistics with concept extraction with respect to an ontology, as well as some initial numerical analysis of the resulting statistics. We should note up front that this is just a pilot project and the computational linguistic method used is really basic, not state of the art. However, even with the simple methods applied, the relationship to formal ontology is relatively novel, and the results appear to be sufficient to provide practical utility in support of a travel application.

Most current travel applications, as opposed to web search, have a limited search structure for the features of hotels. Most search on price and location, as well as a few categories defined by the search provider. A few also expose "amenities" that are self-reported according to standard industry lists. But hotel features that travelers care about can be almost anything stated in natural language. Fine grained search by features would be advantageous. But language is so flexible, some method is also needed to standardize concepts, so that sentiment expressed on the same features can be combined across a large number of reviews.

There are many publications that rate hotels. Each has its own rating approach and scale. If sentiment can be extracted from reviews, those scales can be normalized according to the source data, rather than the reported summary scores.

Sentiment that is linked to concept extraction has the potential to provide a much finer-grained assessment of hotel quality, with respect to the features that each traveler cares about. Not all travelers are the same in their concerns and preferences. More information about the quality of hotel features should considerably improve customer satisfaction.

2 Background: Wordnets

Since Princeton's WordNet (PWN) is well-known, it may be sufficient simply to refer the reader to (Fellbaum, 1998). For the purposes of this paper, it bears mentioning that there are several features of WordNet that make it an essential product to link to.

- PWN is a mature product, having been started over two decades ago (Miller, 1985)
- It is very comprehensive, with over 115,000 word senses, making it the largest wordnet in existence
- It has been free since the project's inception
- It is richly interconnected as a semantic network
- Many other languages have linked their wordnet projects to it manually

3 Background: Suggested Upper Merged Ontology

We had previously mapped all of PWN to a formal ontology (Niles & Pease, 2003), the Suggested Upper Merged Ontology (Niles & Pease, 2001).

Synsets map to a general SUMO term or a term that is directly equivalent to the given synset (Figure 1). New formal terms created for any particular domain will be defined to cover a greater number of equivalence mappings, and the definitions of the new terms will in turn depend upon existing fundamental concepts in SUMO. The process of formalizing definitions will generate feedback as to whether word senses in WN

need to be divided or combined and how the glosses may be clarified. Since many wordnets in other languages are already linked by synset number, this work benefits wordnets in other languages as well.

The Suggested Upper Merged Ontology (SUMO) (Pease, 2011), (Pease&Niles, 2002), (Niles&Pease, 2001) is a freely available, formal ontology of about 1000 terms and 4000 definitional statements. It is provided in a first order logic language called Standard Upper Ontology Knowledge Interchange format (SUO-KIF) (Pease, 2000), and also has a necessarily lossy translation into the OWL semantic web language. It has undergone nine years of development, review by a community of hundreds of people, and application in expert reasoning and linguistics. SUMO has been subjected to formal verification with an automated theorem prover. SUMO has been extended with a number of domain ontologies, which are also public, that together number some 20,000 terms and 80,000 axioms. SUMO has been mapped by hand to the WN lexicon of over 115,000 noun, verb, adjective and adverb senses, which not only acts as a check on coverage and completeness, but also provides a basis for application to natural language understanding tasks. SUMO covers areas of knowledge such as temporal and spatial representation, units and measures, processes, events, actions, and obligations. Domain specific ontologies extend and reuse SUMO in the areas of finance and investment, country almanac information, terrain modeling, distributed computing, endangered languages description, biological viruses, engineering devices, weather and a number of military applications. It is important to note that each of these ontologies employs rules. These formal descriptions make explicit the meaning of each of the terms in the ontology, unlike a simple taxonomy, or controlled keyword list. SUMO is the only formal ontology that has been mapped to all of WN, and the only formal upper ontology that has been extended with a number of domain ontologies that are also open source. SUMO has natural language generation templates and a multi-lingual lexicon that allows statements in SUMO-KIF and SUMO to be expressed in multiple natural languages. These include English, German, Arabic, Czech, Italian, Hindi (Western character set) and Chinese (traditional characters and pinyin).

4 Project Description

We'll first describe the process and then describe the technical details of how it works.

Take for example the following (slightly fictionalized) hotel reviews

Meadowland Resort, Vineland, CA

“In recent years the elegant but unstuffy dining room has won rave reviews, becoming a destination restaurant.“

Crystal Lake Lodge and Resort, CO

“Not to mention it is very expensive and located in a place that doesn't get much sun so it's icy and cold; and the maintenance of roads is terrible in winter.”

The first review is very positive. We extract SUMO concepts from the sentence, and associate them with a positive score. In this case, we assert that the concept of Restaurant has a sentiment of +10. The second review is quite negative. We assert that the concept of Roadway has a -8 sentiment.

We gather reviews for all the hotels we can, extract SUMO concepts and associate them with sentiment scores. We then total the scores for each concept with regard to a particular hotel to create a matrix of hotels and the total sentiment for every concept associated with each hotel.

	Restau-	Break-	Walk-	Fire-	
	rant	fast	ing	place	City
Hotel 1			1		5
Hotel 2		4			
Hotel 3					
Hotel 4	10		10		
Hotel 5	6		6		
Hotel 6				15	11
Hotel 7		1			
Hotel 8	-3				-23

In this section of the resulting matrix with see that Hotel 4 has positive sentiment associate with its breakfast and restaurant. In contrast, Hotel 8 has a very negative sentiment associated with the city (or likely the section of the city) in which it is located.

An open question is how to normalize the sentiment scores. Currently, they are just totaled, which gives greater weight to those cases where there are a large number of reviews. But reviews typically do not mention all the same concepts. One can make the case the frequent mention does legitimately make a statement about the strength of sentiment, since neutral concepts are not likely to be mentioned in reviews.

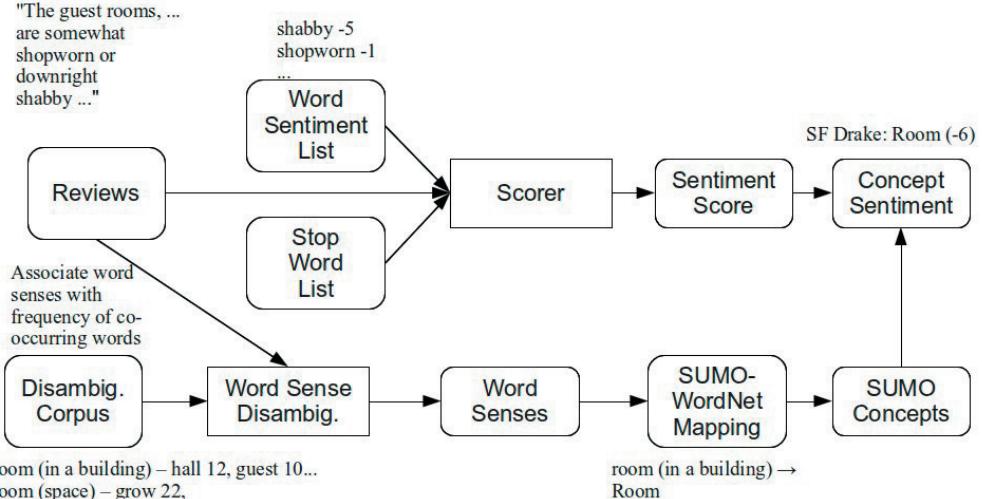


Figure 1: SigmaSentiment data flow

5 Architecture and Data Flow

We now describe all the different elements of the approach as diagrammed in Figure 1. The first step when processing reviews is to determine the sense of each word with respect to WordNet. To accomplish this, we employ WordNet SemCor – a manual markup of word senses in the Brown Corpus. We created a matrix of statistics that associates each word sense with the non-disambiguated words which it co-occurs with. We can then process each sentence, looking at each polysemous word and all other words in the sentence. The sense in SemCor that has the most words in common with the given sentence is the one that “wins” and is selected. This is not a particularly advanced method, and a much larger corpus would be desirable to get better statistical significance, but it is the best that we have found that is also available open source.

For example, consider the word “bed” in the context of the sentence “The bed was so comfortable I had a great night’s sleep.” An amended list of SemCor’s associations between sense and words (also omitting word counts) shows:

Bed	Co-occurring words
sense	
1	air_mattress curtain sleep sleeping_bag slipper
2	compost decayed manure pansy spade spread_over yard
3	dry face homely river tilt

“sleep” is the only word in this simple example that appears in the word lists, and so the highest associated score is with sense 1.

We improve the statistical significance of our word sense discrimination by combining those WordNet senses that map to the same SUMO term. While, SUMO is large, it is not as large as WordNet, and many WordNet senses map to the same SUMO concept. In addition, there has been some discussion that WordNet’s sense may present distinctions that are arguably too fine, and so may not reliably be discriminated in actual usage.

Once we have determined the WordNet sense, then we have a mapping to the appropriate SUMO term.

In parallel with determination of the SUMO terms, we need to calculate a sentiment score for each sentence. The OpinionFinder team has released a file of WordNet senses that have been manually marked with a sentiment score (Wiebe&Mihalcea 2006). Each sense is marked positive or negative, and whether that sentiment is weak or strong.

type	word	POS	polarity
weak	abandon	verb	negative
weak	abate	verb	negative
weak	abdicate	verb	negative
strong	aberration	adj	negative
weak	able	adj	positive

We arbitrarily assign the value of +/-5 to strong sentiment and +/-1 to weak sentiment.

Next we tag every SUMO concept extracted from each sentence with the total sentiment score

for that sentence. All the scores are totaled for all reviews for a given hotel to arrive at a total sentiment for each SUMO concept associated with each hotel.

6 Ontology and Lexicon Development

Since April of 2011, we have been expanding SUMO and the SUMO-WordNet mappings to cover topics in travel and tourism in much greater detail. Significant, new ontologies of Dining, Food and TransportationDetail have been created. Many other ontologies are under development and are available as “beta” ontologies that extend SUMO. These include, Biography, Catalog, Contract, LoyaltyProgram, Pricing and TravelPolicy. Each ontology may have hundreds of terms and formal axioms. Over the same period there have been hundreds of revisions and corrections to existing SUMO ontologies as well. Roughly 2300 SUMO-WordNet mappings have been changed to map to the new more specific concepts now available in SUMO.

Using SUMO terms as the structure to which we attach sentiment scores has several attributes. SUMO is a consistent logical theory, so we are guaranteed that if we follow transitive links, such as subclass, that the results are still inferentially valid. Using a language-independent formal ontology also supports future extensions such as presenting sentiment results in any target language supported by the SUMO language mappings, regardless of the fact that reviews may be processed from English. Lastly, the current work is just the beginning. We plan on moving from concept extraction to statement extraction, and then associating sentiment with entire logical statements, rather than just a mention of a particular concept.

7 Evaluation

We tested our algorithm against a standard test corpus for sentiment analysis (Pang et al 2002). This corpus consisted of roughly 10,000 sentences taken from movie reviews from RottenTomatoes.com using the “fresh” and “rotten” scores supplied by the reviewers as a proxy for a positive or negative sentiment ground truth rating.

We also compared our simple algorithm to OpinionFinder (Wiebe et al 2005), which is a system that assesses subjectivity in text. Because it does not have the same simple goal of just assessing sentiment, comparison is somewhat un-

fair, but it did provide a useful baseline for evaluation.

OpinionFinder does not rank polarity at the level of sentences. We used OpinionFinder’s polarity marks on individual words in the text. We used OpinionFinder as a tool but the final scores at the sentence level are the results of our approach. We sum up the scores that OpinionFinder provides at the word level to give an overall assessment of sentiment at the sentence level.

Pang et al have shown results of up to 80% correctness for sentiment scoring on machine-learning based systems.

% correct	Opinion Finder	Sigma Sentiment
+ reviews	15.04%	64.15%
- reviews	50.78%	49.39%

Figure 2 shows how well SigmaSentiment does with respect to the test corpus of positive reviews, with the scores broken out by sentiment score. It shows the difficulty of determining sentiment when a sentence is mild or ambiguous, but that strong sentiment is relatively easy to determine correctly.

We should note that this problem is quite difficult for humans also. Take for example the fol-

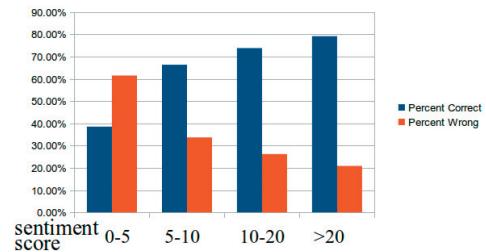


Figure 2: SigmaSentiment incorrect and correct scores on positive reviews

lowing movie review excerpt.

“A disturbing and frighteningly evocative assembly of imagery and hypnotic music composed by Philip Glass.”

Is this review positive or negative? It really depends on whether one finds movies that can be described as “disturbing” to be thought-provoking and therefore interesting and enjoyable, or just unpleasant.

In fact, human raters typically agree about 80% of the time (Wilson et al 2005), where the task was slightly easier in that it allowed raters to agree on “neutral” ratings that correspond to the

most difficult cases where sentiment is not strongly present. As a result, we believe the 80% score should be compared to the portion of our tests performed on sentiment scores above 5 (see Figure 2). A 80% accurate program is doing as well as humans, and our scores of 65%-79% on sentences with strong sentiment may be considered encouraging.

We have also been collecting example sentences from our domain of hotel reviews to calibrate the results we can expect. We place them into three categories. The category of “regular difficulty” contains those sentences with clear sentiment resulting from adjectives. For example,

“The paint of the room is ugly.”

“We had a pleasant stay there.”

The second category is “medium difficulty”, which includes sentences with contrasts or those where some implication or inference is needed. For example,

“The faucet was leaking and making noise whole night.”

“The bathroom is gorgeous but the shower doesn’t work properly.”

In the first sentence it is possible that “leaking” and “noise” could be scored as negative in isolation, although it would be better to know that noise is bad in the context of sleep, which is a primary purpose of a hotel stay. In the second case, some level of analysis is needed to separate the first positive part of the sentence from the second, negative part.

The last category is “most difficult, may need human intelligence.” For example,

“The warm welcome atmosphere disappeared right after I checked into the room.”

“I found that the mattress is no younger than my age although we’re told the rooms were completely renovated last year.”

In these cases, one needs to apply significant common sense knowledge to understand the sentence.

8 Conclusions and Future Work

Language is very flexible, and there are many ways for the algorithm to make the wrong guess at the sentiment associated with a given concept. It appears that given the large volume of reviews we have available, and given that the algorithm gets the sentiment right in a majority of the cases with unambiguous sentiment, that the errors are overwhelmed. More work is needed to see if

there are a statistically significant set cases of errors that can be reliability corrected.

One common error case is in sentences that are divided into positive and negative sentiment, such as the pattern “I liked the bed, but didn’t like the cleanliness of the bathroom.” If we apply the Stanford parser (Klein & Manning 2003) to this sentence, we get the following parse tree:

```
(ROOT
  (S
    (NP (PRP I))
    (VP
      (VP (VBD liked)
        (NP (DT the) (NN bed)))
      (, ,)
      (CC but)
      (VP (VBD did) (RB n't)
        (VP (VB like)
          (NP
            (NP (DT the) (NN cleanliness))
            (PP (IN of)
              (NP (DT the) (NN bathroom))))))))
    (, .)))
```

We then assign sentiment

```
(ROOT
  (S
    (NP (PRP I))
    (VP
      (VP (VBD [liked +5])
        (NP (DT the) (NN bed)))
      (, ,)
      (CC but)
      (VP (VBD did) (RB n't)
        (VP (VB [like +5])
          (NP
            (NP (DT the) (NN [cleanliness +1]))
            (PP (IN of)
              (NP (DT the) (NN bathroom))))))))
    (, .)))
```

And flip the polarity of the sentiment under a negation in the parse tree

```
(ROOT
  (S
    (NP (PRP I))
    (VP
      (VP (VBD [liked +5])
        (NP (DT the) (NN bed)))
      (, ,)
      (CC but)
      (VP (VBD did) (RB n't))
      (VP (VB [like -5])
        (NP
          (NP (DT the) (NN [cleanliness -1]))
          (PP (IN of)
            (NP (DT the) (NN bathroom)))))))
    (. .)))
```

There are of course many exceptions where this approach does not work. However, all we need to do is improve the number of correct interpretations. We will be testing against the movie review corpus to see whether this is the case.

Another area of effort is in capturing user preferences. We cannot expect busy users to go through long lists of concepts, specifying that this or that concept is of particular interest or concern. The existing Rearden Commerce hotel search has the capability of the user to select some concepts such as “best for business” or “good for families” in order to influence the ranking of hotels. The ranking is currently done on the basis of amenities that hotels have self-reported, and a matrix in which company developers have made a judgement that particular amenities are relevant to those categories. While the current approach allows only for the presence or absence of those amenities, sentiment analysis allows us to rate those amenities, so for example, even if a pool is desired by families, a bad or dirty pool should not improve the ranking of a hotel in that category.

Another possibility is to ask users to write a short description of their ideal hotel, and use the same concept extraction process used in the reviews, and then match the preferences of the ideal hotel with those hotels that have positive sentiment for those items.

Acknowledgments

We would like to thank Rearden Commerce for supporting this work, and the OpinionFinder team at U. Pittsburgh for making their work available open source.

References

- Fellbaum, C., (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.
- Klein, D., and Manning, C., (2003). Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.
- Miller, G., (1985) “WordNet: a dictionary browser.” In Proceedings of the First International Conference on Information in Data, University of Waterloo, Waterloo.
- Morato, J., Marzal, M.A., Llorens, J., & Moreiro, J (2004). WordNet Applications. In Proceedings of the Second Global WordNet Conference (GWC-2004). Brno, Czech Republic.
- Niles, I., and Pease, A., (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, Proceedings of the IEEE International Conference on Information and Knowledge Engineering, pp 412-416.
- Niles, I., and Pease, A. (2001). Towards a Standard Upper Ontology. In: Proceedings of FOIS 2001, Ogunquit, Maine, pp. 2-9. See also <http://www.ontologyportal.org>
- Pang, B., Lee, L., and Vaithyanathan, S., (2002) Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP 2002.
- Pease, A., (2003). The Sigma Ontology Development Environment, in Working Notes of the IJCAI-2003 Workshop on Ontology and Distributed Systems. Volume 71 of CEUR Workshop Proceeding series.
- Pease, A., (2011). Ontology: A Practical Guide. Articulate Software Press, Angwin, CA. ISBN 978-1-889455-10-5.
- Wilson, T., Wiebe, J., and Hoffman, P., (2005). "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis,"
- Wiebe, J., Wilson, T., and Cardie, C., (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, volume 39, issue 2-3, pp. 165-210.
- Wiebe, J., Mihalcea, R., (2006). Word Sense and Subjectivity. Joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics. (COLING-ACL 2006).

Linking and Validating Nordic and Baltic Wordnets

- A Multilingual Action in META-NORD

Bolette Sandford Pedersen

University of Copenhagen

Copenhagen, Denmark

bspedersen@hum.ku.dk

Lars Borin

University of Gothenburg

Gothenburg, Sweden

lars.borin@svenska.gu.se

Markus Forsberg

University of Gothenburg

Gothenburg, Sweden

markus.forsberg@gu.se

Krister Lindén

University of Helsinki

Helsinki, Finland

krister.linden@helsinki.fi

Heili Orav

University of Tartu

Tartu, Estonia

heili.orav@ut.ee

Eiríkur Rögnvaldsson

University of Iceland

Reykjavík, Iceland

eirikur@hi.is

Abstract

This project report describes a multilingual wordnet initiative embarked in the META-NORD project and concerned with the validation and pilot linking between Nordic and Baltic wordnets. The builders of these wordnets have applied very different compilation strategies: The Danish, Icelandic and Swedish wordnets are being developed via monolingual dictionaries and corpora and subsequently linked to Princeton WordNet. In contrast, the Finnish and Norwegian wordnets are applying the expand method by translating from Princeton WordNet and the Danish wordnet, DanNet, respectively. The Estonian wordnet was built as part of the EuroWordNet project and by translating the base concepts from English as a first basis for monolingual extension. The aim of the multilingual action is to test the perspective of a *multilingual linking* of the Nordic and Baltic wordnets and via this (pilot) linking to perform a tentative comparison and validation of the wordnets along the measures of *taxonomical structure, coverage, granularity* and *completeness*. Currently, Danish, Finnish, Swedish and Estonian wordnets have been linked to Princeton Core WordNet, thereby providing a common, linked coverage of 5,000 core synsets.

1 What is META-NORD?

META-NORD is an EC project closely related to the META-NET initiative whose very general aim is to foster the technological foundations of a multilingual European information society.¹ More specifically, the META-NORD project aims to establish an open linguistic infrastructure in the Baltic and Nordic countries to serve the needs of the industry and research communities.

META-NORD runs from 2011 to the beginning of 2013 and focuses on 8 European languages - Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian and Swedish which each have less than 10 million speakers. The project aims at assembling, evaluating and linking across languages, and making widely available language resources of different types used by different categories of user communities in academia and industry.

Among these Nordic and Baltic resources are wordnets, which have been developed or are being developed for most of the involved languages. In this paper we investigate the different

¹ For more information on META-NORD and META-NET, see www.meta-net.eu.

nature of these wordnets, and we focus on the perspectives of a linking and evaluation of these in a multilingual context according to certain measures.

2 Multilingual Action on Wordnets

As briefly mentioned, during the last decades, wordnets have been developed for several languages in the Nordic countries including Finnish, Danish, Estonian, Icelandic and Swedish, and just recently a Norwegian wordnet is being initiated on the basis on the Danish wordnet. Of these wordnets, Estonian WordNet is the oldest one since it was built as part of the EuroWordNet project in the 1990s (see Vossen 1999). In contrast, the other three wordnets have been recently initiated; the oldest of them being the Danish wordnet which has been under development since 2005 (cf. Pedersen et al. 2009) and the latest the Norwegian wordnet which is initiated in 2011.

The builders of these wordnets have applied different compilation strategies: Where the Danish, Icelandic and Swedish wordnets are being developed via monolingual dictionaries and corpora and subsequently partially linked to Princeton WordNet; the Finnish and Estonian wordnets have applied the translation method by translating Princeton WordNet into their respective languages for later adjustment.

From the above mentioned different time perspectives and compilation, there was a need for upgrade of several of the wordnet resources to agreed standards, which was thus a preliminary task of this META-NORD action, in particular for Estonian WordNet.

A prerequisite for multilingual use of the resources is that the monolingually based resources are enhanced with regards to either synsets and/or more links to Princeton WordNet. From these links, which primarily constitute the so-called “core synsets” extracted at Princeton University,² pilot cross-lingual resources are derived and further adjusted and validated.

Currently, the linking of the monolingually based wordnets (Icelandic, Swedish and Danish wordnets) to the 5,000 “core synsets” has been completed to ensure common coverage between

all wordnets. A tentative comparison of the resources is planned along the measures of *taxonomical structure, coverage, granularity* and *completeness* (see Section 4).

An additional aim of the multilingual task is to make the relevant wordnets accessible through a uniform web interface.

Wordnets provide semantically-based concept hierarchies for specific languages and are therefore ideal resources to use as a starting point for cross- and multilingual resources; actually they are conceptually better suited than bilingual dictionaries. With such linked resources, cross- and multilingual IR applying semantically-based query expansion becomes feasible. Another possible application for these resources is MT. The hierarchical structure of wordnets ensures that a translation can be found (going up or down in the hierarchy) even if a precise equivalent is not present between the specific languages (this feature is seen in the linking to Princeton Core where some links are constituted by means of the relation eq_has_hyperonym rather than the direct eq_has_synonym).

3 Background of Nordic and Baltic Wordnets

3.1 Estonian Wordnet, EstWN

Estonian wordnet, EstWN was initiated at the University of Tartu in 1990 as a part of the EuroWordNet project. The wordnet was developed by translating the basic concepts from English into Estonian and by building the rest of the wordnet on monolingual grounds. At present it contains more than 45,000 synsets, including nouns, verbs, adjectives and adverbs, as well as some multiword units.

EstWN has been compiled manually but there are some endeavors for automatic additions. For example, a number of words have been derived via suffixes. EstWN includes domain vocabulary from domains such as architecture, agriculture, transportation, and personality traits (Kerner et al 2010).

New updates here? Under the META-NORD project EstWN is being converted to XML-format in compliance with the recently completed KYOTO project. Furthermore, a state of the art editing tool, which can produce XML

² <http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt>

markup, is being developed for further extensions.

3.2 Finnish WordNet, FinnWordNet

FinnWordNet is a recently built wordnet for Finnish developed at the University of Helsinki (cf. Lindén & Carlson 2010). It complies with the structure of Princeton WordNet. It was created by translating all the synsets in Princeton WordNet, and it is open source and contains over 117,000 synsets. After the translation, various things have been done in order to check the quality of the manual translations, e.g. spelling correction, word class consistency correction and some translation correction.

Currently, methods for improving and expanding the content of FinnWordNet are being developed. We have tested methods for finding a location for a new word in the FinnWordNet hierarchy. Since wordnets are structured ontologies, a location for a word can be pinpointed by its relations to other words. Finding a location for a new word means finding a hypernym, a hyponym or a synonym in FinnWordnet. The methods include searching for multiword terms, compound words and using lexico-syntactic patterns. It has also been explored which types of corpora are useful for this task and WikiPedia was found to be valuable in several ways due to its multilingual nature as well as its textual structure.

3.3 Swedish Wordnet, Swesaurus

Swesaurus, a free Swedish wordnet developed at Språkbanken, University of Gothenburg, is constructed by reusing information about lexical-semantic relations in a number of pre-existing freely available lexical resources: SALDO (Borin et al. 2008; Borin and Forsberg 2009), SDB (Järborg 2001), Synlex (Kann and Rosell 2006) and Swedish Wiktionary.

The SALDO resource constitutes the backbone of Swesaurus. SALDO is a large-scale lexical resource providing an inventory of 117k persistent sense identifiers, a morphology of 1.7 MW, and associative semantic relations connecting all senses (somewhat similar to 'evocation' in the WordNet context; Boyd-Graber et al. 2006).

A novel feature of Swesaurus is its fuzzy synsets derived from the graded synonymy relations of Synlex. The recognition of fuzzy synonymy rais-

es many intricate methodological and theoretical questions, e.g., the effect on other lexical-semantic relations, such as hyponymy or meronymy.

As part of the META-NORD project, a linking between Swesaurus and Core WordNet has been completed. The linkage was bootstrapped by using the Lexin basic Swedish-English dictionary. Swedish lemmas in Lexin were automatically linked, in an overgenerating manner, to SALDO sense identifiers, giving us a set of senses for every lemma. The glosses of Core WordNet were subsequently, via Lexin, linked to these sense sets. Core WordNet has 5,000 entries, of which around 89% were covered by Lexin. Furthermore, 23% had a unique link to one SALDO sense, and the remaining an average ambiguity of 4.4 (a rather high ambiguity, but not unexpected for a core vocabulary).

Swesaurus is a part of a larger lexical project, SweFN++, and its development version is published through the lexical infrastructure of SweFN++ on a daily basis. Swesaurus and several other lexical resources are open source available for download and inspection at spraakbanken.gu.se/eng/sblex.

3.4 Danish Wordnet, DanNet

In contrast to most other wordnets, DanNet has been constructed using the so-called merge approach where the wordnet is built on monolingual grounds and thereafter merged with PWN. DanNet is open source and currently contains 65,000 synsets available from www.wordnet.dk in owl/rdf and csv formats (Pedersen et al. 2009). It can be inspected in a browser from www.andreord.dk. The wordnet has been compiled as a collaboration between the University of Copenhagen and the Danish Society for Language and Literature.

Since the starting point of DanNet was a corpus-based, newly completed dictionary of Danish accessible in a machine-readable version with hypernymy information explicitly specified for each sense definition (Den Danske Ordbog), the motivation for the monolingual approach was obvious. Furthermore, the Danish version of the SIMPLE lexicons (cf. Lenci et al. 2001, and for Danish Pedersen & Paggio 2004) has influenced the construction of DanNet in the sense that it includes also qualia information such as the telic (PURPOSE) and the agentive role (ORIGIN).

Qualia roles are encoded in DanNet in terms of relations such as used_for and made_by as well as by means of features such as SEX and CONNOTATION.

3.5 Icelandic Wordnet

Icelandic wordnet is in its early stage of development. It applies the monolingual approach and builds on previous work in the extraction of lexical semantic information from a monolingual dictionary of Icelandic (Nikulásdóttir and Whelpton, 2009; Nikulásdóttir, 2007 Nikulásdóttir & Whelpton, 2010) and seeks to use a mixture of pattern matching and statistical methods for relation extraction, given the promising results from this hybrid methodology in recent years (Cederberg and Widdows, 2003; Cimiano, 2006; Pantel and Pennacchiotti, 2008).

3.6 Norwegian Wordnet(s)

The compiling of a Norwegian wordnet for Norwegian bokmål and Nynorsk is being launched in 2011 by the language initiative Språkbanken and will be developed by the company Kaldera Language Technology. It has been decided to translate from the Danish wordnet, DanNet, and subsequently adjust to Norwegian. The goal is to complete the wordnet(s) in 2013.

4 Methodological Considerations on Linking and Validation

Where the establishment of multilingual resources has a very clear utility value in language technology applications, the purpose of this linking exercise is in fact twofold: Feasibility of cross-lingual linking via the “core synsets” as well as a comparison/validation of the monolingual wordnets. The linking is performed along the 5,000 Princeton Core synsets on the one hand and the selected languages on the other. The primary aim of the task is rather to provide a qualified feasibility study of the perspectives of such a linking at a larger scale and last but not least to give some valuable insights in the very diverse characteristics of the selected wordnets. The main questions to be examined in such a validation are the following:

- **Taxonomical structure:** Do different approaches generally lead to different taxonomical structures of the lexical networks, and can we to some extent define best practice regarding depth of struc-

ture? (For instance, should wordnets generally cover the layman or the expert perspective?)

- **Coverage.** Are frequent concepts in the target language covered well enough when compiling a wordnet via English? And when deducing it from a traditional lexical resource? Can we define a coverage “pain threshold”? These and related issues will be evaluated using corpora and existing core vocabulary lists.
- **Granularity of the described concepts.** Does a specific approach result in many or few sense distinctions (i.e. synonym sets) for each lemma? Is it possible to identify a technology-oriented best practice for sense granularity (i.e. something that corresponds to main senses in traditional lexicography?)
- **Completeness of synonym sets.** Does a given approach bring about many or few semantic relations and/or semantic features per concept? And can a best practice set of semantic relations be established along the validated wordnets?

For illustration of difference in taxonomical characteristics, consider Figure 1 and 2 which show discrepant approaches regarding when to apply a zoological, highly taxonomical perspective and when to apply a simpler, layman approach.

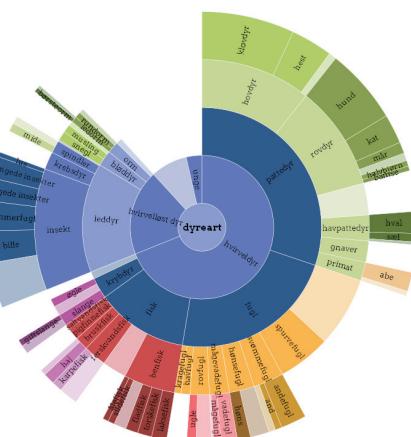


Fig. 1: Taxonomy of animals in DanNet, highly inspired by the zoological taxonomy.

In the case of animals, DanNet adopts a specialist view, where the Icelandic has taken a one-dimensional, layman perspective.³

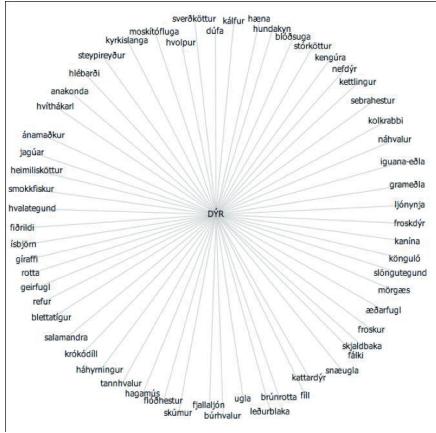


Figure 2: Flat taxonomy of animals in Icelandic Wordnet following a layman approach.

Also if we consider the sets of semantic relations established in the relevant wordnets, we find substantial differences which require further examination. Although the Danish and Swedish wordnets both adopt monolingual approaches, DanNet relates in a stricter way to classical wordnet relations than SALDO/Swesaurus, as is shown in Figure 3 and 4.

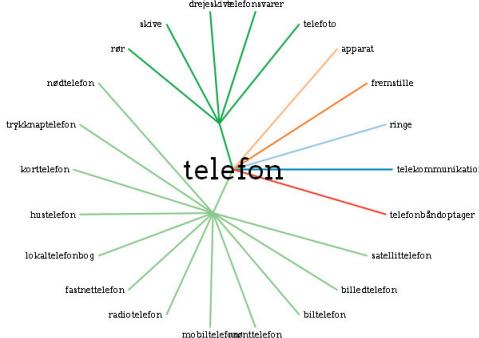


Figure 3: Semantic relations to telephone in DanNet following basically the lines of the Princeton WordNet relations (light green illustrates

has_hyponym, yellow has_hyperonym, dark green has_mero_part, light blue purpose_of etc).

In the Swedish wordnet we find a slightly more associative approach to semantic relations where telephone is furthermore associated to concepts like *samtala* ‘hold a conversation’, *telefonledes* ‘by phone’, *mobiltelefon* ‘mobile phone’.

lex:	telefon
l:	[telefon.e_nn]
fm:	[samtal_e]
fp:	[PRIM]
PRIM:	fingeraukva hørteléfono kohra? pulival ringa telefonautomat telefonera telefonledes telefonur torvald
MIL:	hørteléfono
knapp:	knapptelefon
lokal?	lekaletefon
lyssna:	hörur
mobilt:	mobiltelefon
port:	portmon
radiot:	radiotelefon
viðspg:	vággsíðulefon
absonent:	telefonabsenenti
ansrop:	telefonansrop
apparat:	telefonapparat
avslut:	teleavslut
central:	centralstation
eldedning:	telefondæmning
fingerkvík:	petmjø
førstetailede:	teleførstetailede
forbindelse:	teleforbindelse
kontakt2:	jæk
samtal:	telefonsamtal
signal:	telefonsignal
sladd:	telefonsladd
svev:	telefonsuve
teknisk:	teknisktak
ton:	kopplington
uppfinnerare:	Bell

Fig. 5: Semantic (associative) relations for *telefon* in SALDO.

At the current stage of the project, Danish, Finnish, Swedish and Estonian wordnets have been linked to Princeton Core WordNet, thereby providing a common, linked coverage of the previously mentioned 5,000 core synsets. Next step is to provide a viewer which enables evaluators to see the cross-lingual links in a flexible manner so that validation along the lines described above can be performed in a direct fashion. A possible approach could be the one adapted in the Danish viewer ‘andreord.dk’ were cross-lingual links between DanNet and Princeton Core are shown as direct or indirect alignments.

5 Conclusions

According to the BLARK (Basic Language Resource Kit) scheme, wordnets along with treebanks and other resources, are crucial when building language enabled applications. BLARK lists Computer Assisted Language Learning (CALL), speech input, speech output, dialogue systems, document production, information access and translation applications as dependent of wordnets. The semantic proximity metrics among words and concepts defined by a wordnet are very useful in such applications because in addition to identical concepts, the occurrence of words with similar (more general or more specif-

³ Actually, the Danish wordnet differs internally with regards to layman or expert perspective. Based on The Danish Dictionary, the general approach is that of the layman, but in certain corners of the resource, an expert view has proven dominant.

ic) meanings contribute to measuring the similarity of content.

As has been presented in this paper, most Nordic and Baltic countries are in the fortunate situation where wordnets are already built or are being built right now. However, it is crucial that we continuously adapt them to currents standards and let them undergo cross-lingual comparison and validation in order to ensure that they become of the highest possible quality and usefulness for future, hopefully also multilingual applications. META-NORD provides a unique opportunity for such a validation across languages.

References

- Borin, L., M. Forsberg 2009. All in the family: A comparison of SALDO and WordNet. In: *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*. Odense: NEALT, 7–12.
- Borin, L., M. Forsberg, L. Lönnqvist 2008. The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology. In: Joakim Nilsson, Mats Dahlöf and Beáta Megyesi (red.), *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein. Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia 7*. Uppsala: Uppsala University, 21–32.
- Boyd-Graber, J., C. Fellbaum, D. Osherson, R. Shapire 2006. Adding dense, weighted connections to WordNet. In: *Proceedings of the Global Wordnet Conference 2006*. Brno: Masaryk University, 29–35.
- Cederberg, S. and D. Widdows. 2003. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In *Proceedings of the International Conference on Natural Language Learning (CoNLL)*, pages 111–118.
- Cederberg, S. and D. Widdows. 2003. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In *Proceedings of the International Conference on Natural Language Learning (CoNLL)*, pages 111–118.
- Cimiano, P. 2006. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer.
- Fellbaum, C. (ed). 1998. *WordNet – An Electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts, London, England.
- Kann, Viggo, Magnus Rosell 2006. Free construction of a free Swedish dictionary of synonyms In: *Proceedings of the 15th NODALIDA conference*. Joensuu: University of Eastern Finland, 105–110.
- Kerner, K.; Orav, H. Parm, S. (2010). Growth and Revision of Estonian WordNet. In: Principles, Construction and Application of Multilingual Wordnets. *Proceeding of the 5th Global Wordnet Conference: 5th Global Wordnet Conference; Mumbai, India*; (Ed.) Bhattacharyya, P.; Fellbaum, Ch.; Vossen, P.. Mumbai, India: Narosa Publishing House, 2010, 198 - 202.
- Lindén, K. and L. Carlson. 2010. FinnWordNet – WordNet på finska via översättning [FinnWordNet - WordNet in Finnish via Translation]. *LexicoNordica – Nordic Journal of Lexicography*, 17:119–140.
- Nikulásdóttir, A. B. 2007. *Automatische Extrahierung von semantischen Relationen aus einheimischsprachigen Isländischen Wörterbuch*. MA-Thesis, University of Heidelberg.
- Nikulásdóttir, A. B. and Matthew Whelpton. 2010. Lexicon Acquisition through Noun Clustering. *LexicoNordica* 17:141–161.
- Nikulásdóttir, A. B. and M. Whelpton. 2009. Automatic extraction of semantic relations for less resourced languages. In Bolette Sandford Pedersen, Anna Braasch, Sanni Nimb, and Ruth Vatvedt Fjeld Editors), *Proceedings of the Workshop "Wordnets and other Lexical SemanticResources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies", NODALIDA 2009*. Odense, Denmark: NEALT Proceedings Series Volume 7, pages 1-6. Northern European Association for Language Technology (NEALT), Tartu University Library.
- Pantel, P. and M. Pennacchiotti. 2008. Automatically Harvesting and Ontologizing Semantic Relations. In Paul Buitelaar and Philipp Cimiano, editors, *Ontology Learning and Population: Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*. IOS Press.
- Pedersen, B.S, S. Nimb, J. Asmussen, N. Sørensen, L. Trap-Jensen, H. Lorentzen. 2009. DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation, Computational Linguistics Series*. Volume 43, Issue 3:269-299.
- Vossen, P. (ed). 1999. *EuroWordNet, A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, The Netherlands.

Using WordNet into UKB in a Question Answering System for Basque

Olatz Perez de Viñaspre, Maite Oronoz and Olatz Ansa

University of the Basque Country

Donostia

operezdevina001@ikasle.ehu.es, {maite.oronoz, olatz.ansa}@ehu.es

Abstract

This paper presents the use of semantic information at chunk level in a Question Answering system called *Ihardetsi*. The semantic information has been added through a tool called UKB. For this experiment, UKB uses the Basque WordNet to compute the similarity between the chunks. We use this added information to help *Ihardetsi* to choose the correct answer among all the extracted candidates. Along with the description of the system, we outline its performance presenting an experiment and the obtained results.

1 Introduction

Question answering systems deal with the task of finding a precise and concrete answer for a natural language question on a document collection. These systems use Information Retrieval (IR) and Natural Language Processing (NLP) techniques to understand the question and to extract the answer.

Ihardetsi (Ansa et al., 2009), a Basque question answering system, takes questions written in Basque as input and obtains the results from a corpus written in Basque too. The stable version of *Ihardetsi* incorporates tools and resources developed in the IXA group, such as the morphosyntactic analyzer (Aduriz et al., 1998) and the named entity recognizer and classifier (Fernandez et al., 2011). Nevertheless, we can assume that the use of more syntactic and semantic information in *Ihardetsi*, will probably improve the quality of the obtained answers. Let us see, for instance, the next question in Basque:

“Nor izendatu zuten EEBBtako lehendakari 1944. urtean?” (“Who was appointed president of the US in the year 1944?”)

This question belongs to the *Gold Standard* question bank defined for Basque for the

CLEF2008 conference (Forner et al., 2008). In this bank, the answer given as correct for this question is the following:

“Harry Trumanek Franklin Roosevelt ordezkatu zuen EEBBtako lehendakaritzan 1944. urtean.” (“Harry Truman replaced Franklin Roosevelt in the presidency of the US in 1944.”)

The search of the named entity “EEBB” (“US”), the common noun “lehendakari” (“president”) and the date “1944” separately, does not guarantee that the president to be found by the system will be from the US. For example, searching in Google these three elements, we obtain among others the sentence “The decision of **President** Edwin Barclay (1930-**1944**) to adopt the **US** dollar as the sole legal tender in Liberia...” in which the president is from “Liberia”. The use of chunks, that is noun and verbal phrases, in the question answering system, i.e. “EEBBtako lehendakari” (“president of the US”), would reduce the searching space of the system.

On the other hand, and as we can see in the previous example, sometimes the terms used in the question and in the possible answers, although are not the same (“president of the US” in the question, and “presidency of the US” in the best answer) the terms refer to the same concept. That is one of the reasons why we decided to use the semantic similarity of the chunks to try to improve *Ihardetsi*. The similarity algorithm we use has the Basque WordNet (Pociello et al., 2010) as its base-ontology. As this ontology lacks of named entities we have included some of them with their corresponding synsets to the dictionary used by the algorithm.

As seen in the previous example, the use of shallow syntactic information and semantic information seems to be helpful, so we have integrated more linguistic knowledge in *Ihardetsi*. We have integrated the IXATI chunker (Aduriz et al., 2004) in the analysis chain and we have used a similarity

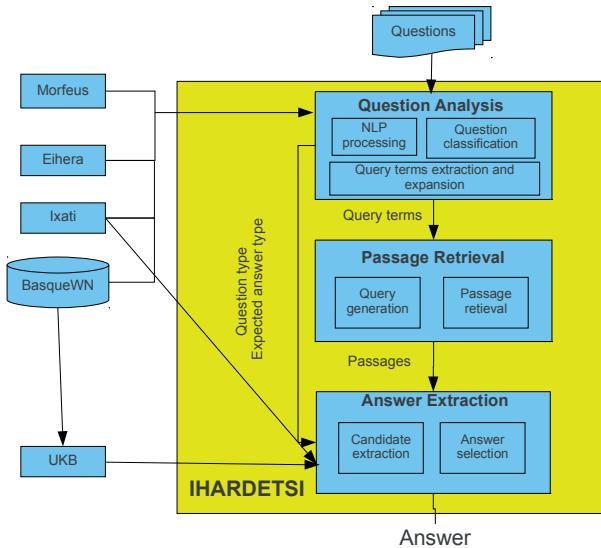


Figure 1: General architecture of the system.

algorithm that is implemented into a tool called UKB (Agirre et al., 2009). The chunks are obtained both in the question and in all the candidate answer-passages.

The remainder of the paper is organized as follows. Section two is devoted to introduce the general architecture of the system. In section three we describe the work done when comparing semantically the chunks from the questions and from the candidate answers. In section four evaluation issues are discussed. Finally, section five contains the conclusions and suggestions for future research.

2 Ihardetsi - A QA System for Basque Language

The principles of versatility and adaptability have guided the development of *Ihardetsi*. It is based on web services and integrated by the SOAP (Simple Object Access Protocol) communication protocol. The linguistic tools previously developed in the IXA group are reused as autonomous web services, and the QA system becomes a client that calls these services when needed. This distributed model allows to parameterize the linguistic tools, and to adjust the behavior of the system.

As it is common in question answering systems, *Ihardetsi* is based on three main modules: the question analysis module, the passage retrieval module and the answer extraction module. Those modules can be seen in the figure 1.

Question Analysis: the main goal of this module is to analyze the question and to generate the information needed for the next tasks. Concretely, a set of search terms is extracted for the passage retrieval module, and the expected answer type along with some lexical and syntactic information is passed to the answer extraction module. Before our contributions, this module used to analyze the questions at morphological level with an analyzer called *Morfus* (Aduriz et al., 1998), and a named entity recognizer called *Eihera* (Fernandez et al., 2011). After the changes described in this paper, the chunker called *Ixati* is added to this module, enriching this way, the question analysis linguistic chain.

Passage Retrieval: basically an information retrieval task is performed, but in this case the retrieved units are passages and not entire documents. This module receives as input the selected query terms and produces a set of queries that are passed to a search engine.

Answer Extraction: in this module two tasks are performed in sequence: the candidate extraction and the answer selection. Basically, the candidate extraction consists of extracting all the candidate answers from the retrieved passages, and the answer selection consists of choosing the best answers among the considered as candidates. The chunker is applied to the candidate answer passages extracted by the stable version of Ihardetsi that uses a kind of “bag of words” technique. For

the work presented in this paper, a re-ranking of the candidate answers is performed using the semantic similarity algorithm from UKB. The number of candidates to be shown could be parameterized but usually five answers are presented to the user.

3 Comparison at chunk level using WordNet

Having applied shallow syntax to the text involved in the QA process, it is possible to compare syntactically the chunks from the question with the according chunks from the candidate answer passages; but also the semantic similarity of the chunks could be measured. Although we have used both syntactic and semantic information to re-rank the answers, we will focus on the semantic area in this paper. The next section describes deeply this work.

3.1 Semantic similarity - UKB similarity

UKB is a collection of programs to perform graph-based Word Sense Disambiguation and lexical similarity/relatedness using a pre-existing knowledge base. It applies the so-called Personalized PageRank on a Lexical Knowledge Base (LKB) to rank the vertices of the LKB and thus it performs disambiguation. The algorithm can also be used to calculate lexical similarity/relatedness of words/sentences (Agirre et al., 2010a) (Agirre and Soroa, 2009).

We took the decision of using UKB according to different reasons: i) it is developed by our same research group, the IXA group; ii) it is language independent as it only needs the semantic knowledge of a language in order to be used, so having a Basque WordNet we can use it for our language; iii) it is free and it is free software as well; iv) it is robust and that is the reason why some analyzers have already integrated it, for example Freeling (Padró et al., 2010).

UKB needs two sources of knowledge that could be extracted from a WordNet to work: on the one hand, a graph containing relations and glosses between concepts, and on the other hand, a dictionary in which word-forms are linked to their corresponding concept. For specific domains as the medical one, other sources as the Unified Medical Language System (UMLS) has been successfully used instead of WordNet in UKB (Agirre et al., 2010b).

Similarity algorithms measure the semantic similarity and relatedness between terms or texts. This concrete algorithm in UKB is able to estimate the similarity measure between two texts, based on the relations of the LKB senses. The method has basically two steps: first, it computes the Personalized PageRank over WordNet separately for each text, producing a probability distribution over WordNet synsets. Then, it compares how similar these two discrete probability distributions are by encoding them as vectors and computing the cosine among the vectors.

When using UKB and WordNet applied to the question answering area, we have found some problems related to the semantic ambiguity of the chunks and to the lack of information in WordNet. These problems will be extensively explained in section 4.

3.2 Procedure to get a weight for each candidate answer

The re-ranking of the candidate answers in *Ihardetsi* is performed by normalizing the weights obtained after the analysis of several syntactic and semantic characteristics. Before explaining this process, in this section we will explain some linguistic phenomena used for the definition of the weights, and then, we will show by means of an example, which features are taken into account for the weight assignment.

We have defined some syntactic patterns at shallow syntax level in order to describe the behavior of some interrogative pronouns in Basque, such as “*Nor*” (“Who”), “*Non*” (“Where”), “*Noiz*” (“When”) and “*Zein*” (“Which”). They all belong to the *factoid* question type, i.e. the question is asking about a simple fact or relationship, and the answer is easily expressed, usually by means of a named entity. We found some interesting phenomena when defining the patterns:

- “*Zein*”. A noun following the pronoun.

The interrogative pronoun “*Zein*” (“Which”) is more complicated than the others. When a noun (or noun phrase) is following the interrogative pronoun (“Which nation is...”) it behaves different, because this noun specifies which the answer should be (a nation in the example). In other words, we know which semantic concept (or synset) are we looking for, so we can find different word forms or instances related to the same concept.

- “Zein”- “Non”/“Noiz”: overlap in the meaning.

Continuing with the “Zein” (“Which”) interrogative pronoun, we have noticed that depending on the meaning of the noun following the pronoun, the pattern is similar to the patterns in other interrogative pronouns. For example, “Zein lekutan dago gailurrik altuena?” (“In which place is the highest peak?”) is equivalent to “Non” (“Where”). In these cases we have added the “Non” patterns to the “Zein” patterns as their meaning is the same. The interrogative pronoun “Noiz” is similar to “Non”, so we added the “Noiz” patterns to the “Zein” patterns as well.

As we have shown, an analysis of the question patterns gives tips to find related concepts in the answers (specific locations and dates) and tips to share patterns among interrogative pronouns.

After having all the chunks tagged in the question and candidate answer passages and having the syntactic patterns properly defined, we perform pattern matching. For the evaluation of the system, we obtain a weight for each element to be compared. These are the elements that are compared: the noun phrase, the verb phrase and the noun (this element is used to evaluate the noun following the interrogative pronoun “Which”).

The following example identifies each of these elements in a question and in one of its candidate answers. In all the cases the lemma of the words is used in order to make easier the deal with the high inflection of the Basque language.

- Question: “Nor izendatu zuten EEBBtako lehendakari 1944. urtean?” (“Who was appointed president of the US in the year 1944?”)
 - Noun phrase: “EEBB lehendakari 1944 urte” (“president of the US in the year 1944”)
 - Verb phrase: “izendatu izan” (“be appointed”)
- Candidate answer: “Harry Trumanek Franklin Roosevelt ordezkatu zuen EEBBtako lehendakaritzan 1944. urtean.” (“Harry Truman replaced Franklin Roosevelt in the presidency of the US in 1944.”)

- Noun phrases: “Harry Truman”, “Franklin Roosevelt” and “EEBB lehendakaritza 1944 urte” (“in the presidency of the US in 1944”)
- Verb phrase: “ordezkatu izan” (“be replaced”)

Following the example, we compare the noun phrase in the question (“EEBB lehendakari 1944 urte”) with all the noun phrases of the candidate answer (“Harry Truman”, “Franklin Roosevelt” and “EEBB lehendakaritza 1944 urte”). Thus, we get the marks given by the similarity script for the three comparisons. We will choose the highest mark as the noun phrase weight to evaluate this candidate answer. In this case, we expect that the chunk “EEBB lehendakaritza 1944 urte” is the semantically closest chunk, and thus, this will be the representative chunk of the candidate answer.

Let us show the marks obtained in this example. In case of the noun phrases “Harry Truman” and “Franklin Roosevelt” the marks obtained from UKB are the same. This is caused by the fact that both are named entities, and they do not appear in our dictionary. As we will explain in the following section, in those cases we get the synset of the entity type, that is, the synset of person. It is surprising that we obtain worse marks comparing the noun phrase “EEBB lehendakaritza 1944 urte” with the question’s chunk. As we have mentioned before, we were expecting this chunk to be best one, but we get the mark 0.0025 over the 0.0040 from the other two. We should explain that the mark 0.0025 seems to be low, but it is a really good mark, considering that the similarity obtained when comparing the chunk with itself (“EEBB lehendakari 1944 urte”) is 0.0037. Anyway, we will take the highest mark, 0.0040, to compare with the other candidate answers. This process is repeated for the verb phrases, and the whole comparison is again applied to the other candidate answers. Thus, we will be able to normalize the weights obtained and according to this normalization to get the best candidate answer.

We can not forget too the cases in which we find a noun following the “Which” interrogative pronoun. In these cases, we have an additional weight: we compare the noun from the question with each one of the concrete answers (not with the passages). For example, for the question “Zein herritan jaio zen Mikel Laboa?” (“In which town was Mikel Laboa born?”) the noun following the

interrogative pronoun is “town”. In this case, we will get the semantic similarity between this noun and the candidate answers. This way we are looking for instances of the concept (“town”), such as “Donostia” or “London”, and we are excluding other kind of entities or nouns, such as “EHU” (an organization) or “research”.

3.3 Problems with WordNet in QA

Clark et al. (2008) expose very clearly the important limitations that WordNet has for supporting textual QA. One of the biggest challenge and in the same proportion important task is the recognition of textual entailment. For this task, we found different semantic knowledge requirements to take into account, such as derivational links, synonyms or world knowledge. Some of those are included in the current WordNet such as synonyms, hypernyms or relations. Other requirements are well oriented for the English WordNet such as the morphosemantic links. In our case, we should follow the path shown on Clark et al. to extend the current Basque WordNet.

Otherwise, named entities have a weighty relevance in QA systems. To treat this lack in WordNet Toral et al. (2008) present a new version of the WordNet: Named Entity WordNet (NEWN). This extension includes named entities extracted from the Wikipedia. Suchanek et al. (2007) developed the YAGO ontology, that is based on WordNet and Wikipedia. Both choices are very interesting for the work presented in this paper but both are offering a solution for the English language. NEWN and YAGO could be translated from English to Basque but as it is a very arduous task, we have found a middle way solution: we have extended the dictionary given to UKB with named entities (person names and location names). The entries added to the dictionary have been extracted from several lists of the Academy of the Basque Language (*Euskaltzaindia*¹). More concretely, we have used toponyms, exonyms and person names in Basque. On the other hand, some named entities identified by Eihera are added to the dictionary following these steps: i) UKB always looks for the target word in the dictionary; if this word is a named entity and it is found in the dictionary, its synset number is used by UKB; ii) if the named entity does not appear in the dictionary, the synset number of the general category obtained by *Ei-*

hera (person, organization or location) is passed to UKB; iii) if a word does not appear neither in the dictionary nor among the entities, a synset number that does not exist is given to the entity, to avoid a crash from UKB. We should continue expanding more the dictionary, for example, with named entities found in the Basque Wikipedia.

Furthermore, we think that by adding acronyms to this dictionary the accuracy of the system will improve. In addition, we will overcome one of the lacks of the Basque WordNet.

4 Experiment and Results

This section describes the results we obtained in our first evaluation of the new version of *Ihardetsi*.

4.1 Experiment

Our corpus for evaluating the system is the “Gold Standard” question bank from CLEF. Those questions were created to evaluate Basque-Basque QA systems. The question bank is composed of 500 questions and their corresponding answers, and we have filtered them just to get the question types we have defined syntactic patterns for. Thus, we only get 63 questions from 500. This subset of the original corpus has been divided into two groups, one for training, and the other one for testing: we get 39 questions for training and 24 for testing.

During the evaluation of the training question bank (39 questions), we noticed that as the question group is small, it is very difficult to draw any conclusion. In addition, some of the questions of the bank were not useful for our system for two main reasons: i) there were problems in the analysis chain, and ii) the stable version of *Ihardetsi* working with a “bag of words” technique returned no results, so the chunking could not be performed. When we start evaluating the training corpus, we realized that, before continuing with the evaluation, a deeper analysis of the re-ranking and the errors in the analysis chain was necessary. Thus, we decided not to evaluate the test corpus. The results in table 1 correspond to the training question bank.

As far as the corpus is concerned, it is composed of all the documents of the *Euskaldunon Egunkaria* (a newspaper wholly written in Basque language) from the years 2000 to 2002, with the amount of 24 million up words (in total). Additionally, the corpus has a version of the Basque Wikipedia from the year 2006 with 1.5 million up

¹<http://www.euskaltzaindia.net>

words.

4.2 Results

As mentioned in section 2, *Ihardetsi* usually shows the best five candidate answers. In the evaluation we decided that it is very interesting to measure if the correct answer is close to the first position and in which position it is.

Table 1 shows the results obtained when testing the following systems: *Ihardetsi* in the original/stable mode (*Ihard*), and *Ihardetsi* using semantic information (*Ihard+S*).

	Ihard	Ihard+S
Correct answer (first candidate)	7	8
Answer in the first five candidates	8	5
Not in the first five candidates	8	10
Total	23	23

Table 1: Results.

As shown in the results, taking into account the few data we deal with, no general conclusions could be reached, but the overall impression is good. The correct answers improve when *Ihardetsi* uses semantic information.

If we go into details, it could be surprising to see that the semantic contribution obtains more correct answers out from the first five candidates. We think that this is caused by the behaviour of UKB. We think that UKB gives higher marks to the less ambiguous chunks and lower marks to those in which the words inside the chunk have a high ambiguity. Usually, in question answering systems the answers have among others named entities. In our system it is very important to analyze if they must be tagged in a general way, giving them an unique general synset (e.g. “location” for “Barcelona”) or concretely assigning them all the possible synsets they have (e.g. “Autonomous city” synset 08524735 and “Province” synset 08654360n for “Barcelona”). This decision could change the obtained results. On the other hand, the lack of named entities affects the results too. As UKB works with the dictionary extracted from the Basque Wordnet an it has very few named entities (among others the ones added using lists as described in section 3.3), we lose accuracy.

We find interesting to go step by step analyzing each phase of the linguistic analysis chain to try to understand the results. In the same way, the reader will be able to understand the problems of the system and how these barriers

could be broken down in order to obtain better results. We will explain those steps with the paper-example: “Nor izendatu zuten EEBBetako lehendakari 1944. urtean?” (“Who was appointed president of the United States in the year 1944?”).

Our module takes two main input files: the analysis of the question and the analysis of all the candidate answer-passages. In the first step we generate a file that contains all the information we need from the question:

```
Question structure: nor VP NP ?
Answer structure: [ENTI_PER] {VP} NP
Noun phrase: eebb lehendakari 1944. urte
(president United States year 1945)
Verb phrase : izendatu edun (appoint)
```

```
UKB format
Noun phrase: EEBB#n#1#0
lehendakari#n#2#0 urte#n#3#0
Verb phrase: izendatu#v#1#0 edun#v#2#0
```

In the second step we extract the information from the candidate answers analysis and we generate a similar file with all the information about all the candidate answers. The similarity algorithm assigns weights to the chunks in the way explained in section 3.2. The semantic ambiguity of the chunks and the numbers and specificity of the synset assigned to the named entities change the results.

Otherwise, we lose some information when we translate the chunks into the UKB format. This is the case of the numbers (i.e. the year “1944”). UKB just works with nouns, verbs, adjectives and adverbs, and as there is not place for numbers, they are tagged as determiners. Although usually numbers are not important for lexical semantics, they are significant in question answering systems.

As a consequence of these problems, the results of our module are not better than the results obtained using the original *Ihardetsi*. Anyway, as we have been able to identify the problems, we will be able to correct them.

5 Conclusions and Future work

In this paper we have presented important contributions to *Ihardetsi*, a Question Answering system for Basque, adding semantic information to *Ihardetsi* by means of a chunker and an algorithm that performs semantic similarity. We have exposed the use of UKB and WordNet for QA, and we have listed some problems related to WordNet in QA, finding a middle way solution to the lack of named entities.

In order to extract more concrete conclusions about the behavior of UKB in *Ihardetsi*, we need a bigger evaluation question bank. The number of valid questions will increase improving the linguistic analysis chain. A newer version of the linguistic tools is ready to be integrated in *Ihardetsi* giving us the chance to improve the results of the system.

The similarity script output has been used as given, but we think that it is necessary to study the impact of the terms ambiguity in the chunk and perhaps these results need to be tuned. Due to the relevance of the named entities in QA systems, we need to expand the dictionary for example with named entities found in the Basque Wikipedia or using other additional ontologies.

The results obtained in the experiment give us a promising way to research. In our opinion, the use of semantic information is very interesting not only for the re-ranking of the candidate answers, but also to help in the candidate extraction task.

Acknowledgments

This research is supported by the Ministry of Science and Innovation of the Spanish Government (KNOW2 (TIN2009-14715-C04-01)) and the Basque Government (IT344-10 and Berbatek (IE09-262)).

References

- Itziar Aduriz, Eneko Agirre, Izaskun Aldezabal, Iñaki Alegria, Olatz Ansa, Xabier Arregi, Mari Arriola Jose, Xabier Artola, Arantza Díaz de Ilarrazza, Nerea Ezeiza, Koldo Gojenola, Montserrat Maritxalar, Maite Oronoz, Kepa Sarasola, Aitor Soroa, Ruben Urizar, and Miriam Urkia. 1998. A Framework for the Automatic Processing of Basque. In *Proceedings of Workshop on Lexical Resources for Minority Languages. First LREC Conference. Granada. 1998*.
- Itziar Aduriz, Maxux Aranzabe, Jose Mari Arriola, Arantza Díaz de Ilarrazza, Koldo Gojenola, Maite Oronoz, and Larraitz Uria. 2004. A Cascaded Syntactic Analyser for Basque. In *Computational Linguistics and Intelligent Text Processing. 2945 pg. 124-135. ISBN 3-540-21006-7 LNCS Series. Springer Verlag. Berlin. 2004*.
- Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009). pp. 33-41. ISBN: ISBN 978-1-932432-16-9*.
- Eneko Agirre, Aitor Soroa, Enrique Alfonseca, Keith Hall, Jana Kraválova, and Marius Pasca. 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proceedings of NAACL-HLT. Pages 19-27. Boulder, Colorado. ISBN: 978-1-932432-41-1*.
- Eneko Agirre, German Rigau, Aitor Soroa, and Montse Cuadros. 2010a. Exploring Knowledge Bases for Similarity. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC10). European Language Resources Association (ELRA). ISBN: 2-9517408-6-7. Pages 373-377*.
- Eneko Agirre, Aitor Soroa, and Mark Stevenson. 2010b. Graph-based Word Sense Disambiguation of Biomedical Documents. *Bioinformatics*, 26:2889–2896.
- Olatz Ansa, Xabier Arregi, Arantxa Otegi, and Ander Soraluze. 2009. Ihardetsi: A Basque Question Answering System at QA@CLEF 2008. In *Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008. Lecture Notes in Computer Science, Vol. 5706/2009, pp. 369-376. Springer Berlin / Heidelberg. ISSN 0302-9743 ISBN 978-3-642-04446-5*.
- Peter Clark, Christiane Fellbaum, and Jerry Hobbs. 2008. Using and Extending WordNet to Support Question-Answering. In *Proceedings of the 4th Global WordNet Conference, Szeged, Hungary*.
- Izaskun Fernandez, Iñaki Alegria, and Nerea Ezeiza. 2011. Semantic Relatedness for Named Entity Disambiguation using a small Wikipedia. In *TSD 2011: 14th Conference on Text, Speech and Dialogue*.
- Pamela Forner, Anselmo Penas, Eneko Agirre, Iñaki Alegria, Corina Forascu, Nicolas Moreau, Petya Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu, Richard Sutcliffe, and Erik Tjong Kim Sang. 2008. Overview of the CLEF 2008 Multilingual Question Answering Track. In *Working Notes of the Cross-Lingual Evaluation Forum, Aarhus, Denmark. ISBN 2-912335-43-4, ISSN 1818-8044*.
- Llus Padró, Samuel Reese, Eneko Agirre, and Aitor Soroa. 2010. Semantic Services In Freeling 2.1: WordNet and UKB. In Piek Vossen eds. P. Bhattacharyya, Christiane Fellbaum, editor, *Principles, Construction and Application of Multilingual Wordnets: Proceedings of the 5th Global Wordnet Conference. ISBN: 978-81-8487-083-1*. Narosa publishing house.
- Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2010. Methodology and Construction of the Basque WordNet. In *Language Resources and Evaluation. Springer. ISSN 1574-020X*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Antonio Toral, Rafael, Munoz, and Monica Monachini. 2008. Named Entity Wordnet. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC08)*, Marrakech, Morocco.

Automated Mapping of Polish Proper Names on plWordNet Hypernymy Structure

Roman Kurc

Wrocław University
of Technology, Poland

{roman.kurc,maciej.piasecki,agnieszka.indyka-piasecka}@pwr.wroc.pl

Maciej Piasecki

Wrocław University
of Technology, Poland

Agnieszka Indyka-Piasecka

Wrocław University
of Technology, Poland

Abstract

The paper presents an algorithm of automated mapping of Proper Names onto plWordNet (a Polish wordnet) synsets. The algorithm is a modification of Algorithm of Activation-area Attachment used for wordnet expansion. Evidence provided by different knowledge sources is analysed on the level of lexical units and their location in the wordnet structure. Influence of each knowledge source on the local context can be tuned to its characteristics. The algorithm was applied to a large set of Polish Proper Names and a set of heterogeneous knowledge sources including sources extracted from the text corpus, as well as, acquired from the meta-data description.

1 Motivation and Related Works

Proper Names (PNs) are the most numerous class of the natural language lexical units (LUs), but only partially described in semantic lexicons, as the class of Proper Names is very dynamically changing. PNs are included in many wordnets, e.g. Princeton WordNet (Fellbaum, 1998), however in small numbers, but not in plWordNet in general. The only exception are PNs that are linked by derivational relations to common nouns also described in plWordNet. Instead, we aim at constructing a kind of open lexicon of PNs which can be expanded on demand, mainly on the basis of large corpora, and whose elements are mapped to wordnet synsets by *instance-type* relation (Miller and Hristea, 2006).

Identification and classification of PNs has been intensively researched for many years as a part of Named Entity Recognition (NER). However, PNs are typically classified in NER into a limited number of classes. Mann

(Mann, 2002) proposed a method of automated construction of PN ontology. Sundheim et al. (Sundheim et al., 2006) studied linking gazetteer elements to WordNet, but only geographical terms. In (de Loupy et al., 2004) and (Toral and Monachini, 2008) the idea of building a thesaurus of PNs on the basis of WordNet was proposed. de Loupy added 130675 PNs “founded on several knowledge bases” to the wordnet, but only to 55 synsets.

Toral and Monachini (2008) presented an approach to expanding of WordNet with Named Entities (NEs) automatically extracted from the English Wikipedia¹. The proposed process consisted of two phases. In the first phase, each WordNet monosemous noun synset is linked to a Wikipedia category only if a synset member is identical to one of the category words. During the second phase, NEs described in Wikipedia by the mapped categories are added to WordNet as new synsets linked by *type/instance* to synsets identified by the mapping. Good results were reported, but it must be noticed that the key problem of disambiguating polysemous associations between the categories and synsets was almost neglected, as well as, the problem of an extensive usage of other knowledge sources than Wikipedia. Ruiz-Casado et al. (2005) proposed a method for automated mapping Simple English Wikipedia articles describing PNs onto WordNet. However, the mapping depends on the vector-based similarity of the synsets’ glosses and the article text contents, while *plWordNet* does not contain glosses.

Mapping of PNs onto a wordnet is very similar to the problem of placing new words in the appropriate locations of the wordnet hypernymy structure. Several algorithms of automated wordnet expansion have been proposed, e.g., (Snow et al., 2006; Piasecki et al., 2009a).

Our goal is to construct a method for auto-

¹<http://en.wikipedia.org>

mated mapping PNs onto wordnet synsets by *type/instance* relation and to create a semantic lexicon of PNs providing their fine grained semantic description. Ambiguous PNs can be linked to several synsets. The algorithm should be based mainly on the analysis of large plain text corpora, and only supported by knowledge extracted from collection of structured documents like Wikipedia.

2 Sources of lexical knowledge

The proposed mapping method is based on Activation Area Attachment algorithm (AAA) (Piasecki et al., 2009b; Piasecki et al., 2009a) for wordnet expansion. AAA works with varied heterogeneous partial knowledge sources (KS) extracted from text corpora and Wikipedia (Piasecki et al., 2011a). KSs are represented as sets of lemma pairs: $\langle x, y \rangle$ such that x is a PN (not described in *plWordNet*, possibly multi-word) and y (nominal lemma from *plWordNet*) is semantically related to x according to the given extraction method and the corpora analysed. We used a joint corpus of about 1.2 billion of tokens, cf (Piasecki et al., 2011a). 119820 PNs described in the Polish Wikipedia² were used in the experiments.

Several heterogeneous KSs were used in experiments presented here of three main types: *Distributional Semantics, pattern-based* and meta-data based. Concerning the first type, a Measure of Semantic Relatedness (MSR) was extracted from the joint corpus by applying the algorithm described in (Piasecki et al., 2007). The MSR was built for PNs and 33577 nouns from *plWordNet* 1.4, but only for those that occur ≥ 50 times and are described by ≥ 10 different lexico-syntactic features. *Two KSs* were produced on the basis of the MSR: the set of the $k = 20$ lemmas most semantically related to an PN and this set restricted to *bidirectional semantic relatedness* between an PN and a wordnet lemma.

Three handwritten lexico-syntactic patterns designed for the extraction of hypernymic and synonymous word pairs, see (Piasecki et al., 2009b) were used here too. In addition, we used *seven manually written patterns (structural patterns)* proposed in (Piasecki et al., 2011a) to be applied to the initial parts of the

Wikipedia articles and extract instance/type pairs. All manually written patterns were activated only for the PNs from the assumed list.

Especially for PN mapping, we also extracted information from *Wikipedia categories* attached to PN-related articles. However, the categories are often too long and too specific to match exactly any wordnet lemma or do not represent a type but stay in a different conceptual relation to the described term, e.g. meronymy and location *Ciotcza; gmina Abramów* ‘Ciotcza [meronym]; commune Abramów [holonym]’. Moreover, categories are mostly in the plural and must be transformed to the singular form first. Contrary to (Toral and Monachini, 2008) we use all categories including those of polysemous matching. From each category we extracted all one-word and multi-word nominals in the nominative case as potential syntactic heads of the category. Concerning multi-word lemmas, only those described structurally in a dictionary created for *plWordNet* were taken into account (6582 MWE categories out of 168678 categories in general, 204 unique MWE categories out of 3325 categories) 168 678 pairs: PN – noun were extracted. Pairs PN – PN were not recorded as they express different types of relations than instance-type, e.g. meronymy.

Finally two more meta-data based KSs were acquired. Firstly, we extracted nominals occurring in *brackets in term names* (PNS) – the bracket expressions represent mostly class names, e.g. *Casablanca (film)* ‘movie’. We obtained in total 1202 unique nominals, 842 of them were MWE. This source consisted of 5628 pairs and it described 3456 of our PNs. Secondly, *coarse-grained classes* assigned to many PNs in the gazetteer of (Marcinčuk and Piasecki, 2011) were used.

3 Modified Algorithm of Activation-area Attachment

AAA works in two phases: *Phase I* in which all synsets that fit semantically a new lemma (here: a PN) are found and *Phase II* – the found synsets are grouped into connected subgraphs – *activation areas*, cf details in (Piasecki et al., 2009b; Piasecki et al., 2009a; Piasecki et al., 2011b). Properties of KSs col-

²<http://pl.wikipedia.org>

lected for PNs initiated changes in *Phase I*, leading to a new AAA version called *Lexical-unit-based AAA* (LAAA), in which *Phase II* is unchanged.

During *Phase I*, for a PN (to be mapped) x :

1. Semantic fit between x and each wordnet lemma y is calculated as a sum over weights assigned to KSs including $\langle x, y \rangle$
– weights correspond to the manually assessed precision of the sources.
2. For each synset S $score(x, S)$ is calculated on the basis of the sum of semantic fit of x to the lemmas of S and fit of x to lemmas included in synsets in the local context of S (i.e. synsets linked to S by short paths in the graph of wordnet relations).

Step 1 is kept unchanged from AAA, see (Piasecki et al., 2011b). However, MSR provides description for only some PNs, i.e. 1088 out of 119820 – contrary to wordnet expansion, where MSR was defined for almost all lemmas.

As all KSs express intrinsically some error, in Step 2 the semantic fit of a new lemma x to a synset S is collected not only from the fit of x to lemmas of S but also from the lemmas in synsets accessible from S via relation links, cf (Piasecki et al., 2009b; Piasecki et al., 2009a). Contextual fit is processed by selected types of transformations sensitive to the distance and types of links in the path from S to synsets in neighbourhood. In AAA, a typical value for the maximal path length (context size) was 2.

Meta-data based KSs provide PN classification. Such classes very often correspond to remote hypernyms of those PNs, not their direct types. In order to utilise this long distances association we introduced in LAAA separate *tracing* of fit from different KSs – for some of them much longer paths are allowed. The calculation of the synset *score*, combining local and contextual fit, is done according to Equation (1) (slightly simplified, as the mechanisms of *conduction* and *impedance*, discussed below, are neglected), where K is the set of KS labels.

In LAAA, for each KS we calculate its contextual influence and apply distance sensitive weighting to it separately. In the case of PN mapping, two groups of KSs were introduced: corpus-based and meta-data based. For the first group the semantic fit is collected from the

distance of at most 2 links and the weight is $1/(2d)$, where d is the path length. For meta-data based KSs the size of the context is limited to 1, except pure hyponymy paths that are not limited, i.e. data pertaining to a synset S can influence all its direct and indirect hyponyms. The distance weight first goes down with the increasing distance but later gets stable on the level of 0.25. Thus, meta-data based fit marks whole hypernymic subtrees.

Fit can be ‘transported’ along paths consisting of different relation links. Link semantics influences feasibility of relating contextual fit via the given link to the given synset, e.g. all information related to a lexical unit can be fully related to its synonym, almost completely to its hyponym, but this transfer is questionable in the case of its antonym. In (Piasecki et al., 2011b) the notions of: *conduction* (simplifying, link ability to transfer hypernymy-related fit) and *impedance* (simplifying, extent to which fit can be transferred across borders between different relations) are preserved in LAAA together with modelling them with the help of weight functions decreasing the value of the transferred fit, cf (Piasecki et al., 2011b).

The difference between strong and weak fit to a synset defined in AAA is not used in LAAA application to PN mapping and, thus, not presented in the description of *Phase I*. During *Phase II*³ first for a PN x continuous areas (connected subgraphs) in the hypernymy graph are identified such that each synset in an area is assigned non-zero score to the given PN. For each area a synset with the local maximum of the score is identified. All subgraphs with the score above some threshold are preserved and returned as descriptions of the possible types of x .

4 Evaluation

As Polish PNs are neither described in *plWordNet* (with some exceptions), nor in any available semantic lexicon, we assumed manual evaluation as the basis. Algorithms’ precision in selecting the appropriate synsets and recall measured in terms of the PN senses covered were targeted. Evaluation was performed using resources described in Sec. 2 and *plWord-*

³A simplified description is given here, all details can be found in (Piasecki et al., 2009b).

$$score(x, S) = \sum_{k \in K} [\sum_{y \in S} fit(x, y, k) + \sum_{S' : cntx(S, S')} [\sum_{y' \in S'} f_w(path(S, S'), k) fit(x, y', k)]] \quad (1)$$

Cont.	T [%]	K [%]	H	total[%]
2	63.25	5.51	8.93	77.68
1	66.76	5.03	8.36	80.15

Table 1: Mapping PNs on *plWordNet* based on AAA and Wikipedia categories.

Net 1.4 (PNs included in *plWordNet* 1.4 were removed from the testing data). In order to analyse the combination of heterogeneous KSs only 119820 PNs covered by Polish Wikipedia (20.10.2010 dump) were used. Moreover, in order to decrease the computational power required for the extraction of KSs (especially for MSR), all experiments were performed on a randomly selected test set of 10000 PNs.

In the first (baseline) experiment AAA and Wikipedia categories as the only KS were used. The results for two sizes of the local context: 1 and 2 are presented in Tab. 1. For each PN up to 5 top-scored attachments (pairs: PN-synset) were generated. For each version a sample of 1050 attachments were randomly selected, that included attachments of all 5 ranks. We assessed manually the sample and divided the results into four classes: *correct type* (marked T), *super type* (a non-direct hypernym) (H), a *co-hyponym* (K) – the suggested attachment and the PN share a close hypernym (1-2 links up the hypernymy structure.) In both cases the obtained results are much worse than 93.02% precision reported in (Toral and Monachini, 2008), however, they focused on PNs with categories corresponding to monosemous lemmas in the wordnet, in our case we performed mapping for all PNs from the test set. Moreover we could not leverage on PNs described in the wordnet as *plWordNet* includes only a small number of PNs.

Wikipedia categories corresponding to polysemous *plWordNet* synsets activate several synsets per PN. Most of them are located in the upper parts of the hypernymy hierarchy and when the context is set to 2 or 1, most of the activated synsets are presented as possible attachments. Many of them describe senses unrelated to analysed PNs. The category-

based KS is too sparse in terms of the hypernymic graph topology. A little surprisingly, the smaller context resulted in slightly better results, as the categories classify PNs and the larger context causes that the classes are wrongly extended to hypernyms.

In the second experiment LAAA and all KSs (Sec. 2) were applied. Contrary to the wordnet expansion experience, cf (Piasecki et al., 2009b), MSR-based KSs became sparse and lost the status of the basic sources. Two LAAA settings were tested: all KSs vs MSR-based KSs excluded. A sample of 500 PNs was randomly selected all 5 top scored proposals were evaluated manually. The precision is presented in Tab. 2: ALL – average precision for all suggestions, FIRST – precision for only one top scored suggestion, and >ONE – percentage of PNs with at least one good suggestion is among the 5 top scored.

Example: George Weah – *<polityk ‘politician’>*; George Weah – *<uczestnik ‘participant’>*; George Weah – *<gracz ‘player’, pilkarz ‘football player’>*.

It is worth to emphasise that the experiments were performed for PNs described by either monosemous or polysemous categories. Results: based only on categories (Tab. 1) and on all KSs (Tab. 2) are close, but in the latter case all 5 suggestions were evaluated per PN. Many PNs do not have 5 different referents, so the overall precision is decreased. LAAA mostly do not define a clear cut between ‘true’ suggestions and the rest. As regards the top scored suggestions the precision is good and shows LAAA ability to disambiguate PN type on the basis of sparse KSs and the wordnet structure. The >ONE class presents LAAA as a tool supporting manual work on the PN semantic lexicon construction: the vast majority of short suggestion lists include proper types. For many PNs MSR produces results of lower accuracy, even with the high frequency threshold. Results obtained without MSR-based sources are better. However, we observed many cases in which MSR helped to disambiguate. A version of PN-dedicated MSR

Top scored synsets only				
Sources	T [%]	K [%]	H	total[%]
ALL	63.3	3.9	10.4	77.6
No MSR	65.3	6.3	7.1	78.7
Only first suggestion				
FIRST	80	1	5	86
No MSR	81.6	4.8	3.6	90
Correct suggestion among the top 5				
>ONE	96	0.8	2.4	99.2
No MSR	92.4	3.6	1.6	97.6
All top 5				

Table 2: Mapping PNs on plWordNet based on LAAA algorithm and heterogeneous KSSs. (No MSR – MSR was excluded)

extraction algorithm is required.

5 Further Research

A wordnet seems to be a good basis for PN semantic description, but due to large PN quantities an automated method is needed. AAA – a wordnet expansion algorithm – appeared to be a good starting point, however, significant extensions were required due to the different characteristics of PN text distribution and KSSs extracted for them. In fact a generalised version of AAA was proposed and the ‘old’ AAA can be perceived only as a specific setting of the algorithm. Limited, sparse KSSs extracted from the corpus for PNs must be supplemented by other possible sources of information like mining definition-like descriptions of PNs and meta-data in a way sensitive to the specific text genre or even text type. A dedicated version of MSR extraction method must be found in order to make this source applicable for PNs. The achieved accuracy for the top-scored suggestions is on a good level, however still a method for selecting only high reliability suggestions and their appropriate number for a given PN is required.

Acknowledgments Financed by the Polish National Centre for Research and Development project SyNaT.

References

- Claude de Loupy, Eric Crestan, and Elise Lemaire. 2004. Proper nouns thesaurus for document retrieval and question answering. *Atelier Question-R` posse, e TALN*.
- Christian Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- Gideon S. Mann. 2002. Fine-grained proper noun ontologies for question answering. In *Proc. of the 2002 workshop on Building and using semantic networks - Vol. 11*, SEMANET’02, pages 1–7, Stroudsburg, PA, USA. ACL.
- Michał Marciniuk and Maciej Piasecki. 2011. Statistical proper name recognition in polish economic texts. *Control and Cybernetics*, 40(2):1–26.
- George A. Miller and Florentina Hristea. 2006. WordNet nouns: Classes and instances. *Computational Linguistics*, 32(1):1–3.
- Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2007. Automatic selection of heterogeneous syntactic features in semantic similarity of Polish nouns. In *Proc. TSD 2007 Conf.*, volume 4629 of *LNAI*. Springer.
- Maciej Piasecki, Bartosz Broda, Maria Głabska, Michał Marciniuk, and Stan Szpakowicz. 2009a. Semi-automatic expansion of polish wordnet based on activation-area attachment. In *Recent Advances in Intelligent Information Systems*, pages 247–260. EXIT.
- Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009b. *A Wordnet from the Ground Up*. OW PWr, Wrocław.
- Maciej Piasecki, Agnieszka Indyka-Piasecka, and Roman Kurc. 2011a. Linguistically informed mining lexical semantic relations from Wikipedia structure. In *Proc. of The 2nd ACIDS*, LNAI. Springer.
- Maciej Piasecki, Roman Kurc, and Bartosz Broda. 2011b. Heterogeneous knowledge sources in graph-based expansion of the polish wordnet. In *Proceedings of The 2nd Asian Conference on Intelligent Information and Database Systems*, number 6591 in *LNAI*. Springer.
- Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In *Advances in Web Intelligence*, volume 3528 of *LNCs*, pages 380–386. Springer.
- Rion Snow, Dan Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogeneous evidence. In *COLING’06*.
- Beth M. Sundheim, Scott Mardis, and John Burger. 2006. Gazetteer linkage to wordnet. *Proc. of the III IWC*.
- Rafael Muñoz Antonio Toral and Monica Monachini. 2008. Named entity wordnet. In ELRA, editor, *Proc. of the VI LREC’08*, Marrakech, Morocco, V.

Automated Generation of Derivative Relations in the Wordnet Expansion Perspective

Maciej Piasecki

Wrocław University
of Technology, Poland

{maciej.piasecki, radoslaw.ramocki, marek.maziarz}@pwr.wroc.pl

Radosław Ramocki

Wrocław University
of Technology, Poland

Marek Maziarz

Wrocław University
of Technology, Poland

Abstract

We present a machine learning approach to the generation of derivative relations. Instances of derivational relations described in a wordnet are used in the bootstrapping approach to build an analyser of derivational relations. plWordNet derivational relations are presented and the planned semi-automatic wordnet expansion with derivational relations is discussed. Limits to which form-based markers can encode semantic distinctions are analysed and a model of semantic post-filtering of the generated derivational relation instances is presented.

1 Introduction

Derivational relations occur in many languages and mostly encode certain lexico-semantic relations, e.g. diminutives. They are present in many wordnets. Due to their regular character and productivity there are attempts to automate wordnet expansion with respect to derivational relations. In existing approaches lacking derivational links are automatically generated by extended morphological analysers, which, however, are based on hand-crafted rules. For many languages, including Polish tools of this type and coverage do not exist.

Our idea is to remove this external dependency and to enclose the automated expansion of the wordnet derivational part in a kind of bootstrapping scheme: start with a handful of instances of derivational relations added manually to a wordnet, train a generator of derivatives and use it to boost the wordnet expansion. The system should be open for different relations and trained on wordnet data. the system should not only generate pairs of words as associated by a formal derivative relation but should also identify a semantic relation that is expressed by a derivational pair.

2 Derivation, Automation and Wordnets

Two problems need to be solved in order to fully automate wordnet expansion with derivational relations: a method must be defined and semi-

automated construction of a generator of derivational links must be provided.

2.1 Automated wordnet expansion

A large scale process was run for Czech WordNet (Pala and Hlaváčková, 2007). 10 “main regular derivational relations” were covered. Relation instances were generated on the basis of the hand-crafted description of the inflection derivational paradigms implemented in an expanded version of the morphological analyser. The results were manually corrected and non-derivational lemmas were manually deleted. As the applied analyser do not support “changes in stem” (alternations) appropriate modifications had to be introduced manually. Derivational relations link lemmas (“literals”) in Czech WordNet, i.e. if a lemma has more than one lexical unit (LU), the lexico-semantic derivational relation is implicitly extended to all LUs pairs pertaining to it. This is in contrast to the semantic ambiguity of suffixes noted in (Fellbaum et al., 2009).

The idea of a semi-automated expansion of the wordnet derivational relations was also discussed in (Koeva et al., 2008) for the Serbian and Bulgarian. In both wordnets derivational relations link synsets, i.e. lemma pairs are expanded on synsets to which they belong. So the semantics of these relations seems to be expanded beyond the associations encoded by derivational pairs of LUs. Our approach presented in Sec. 3 goes in between the distinction: a relation on lemmas vs synsets.

2.2 Generating derivative relations

Works dedicated to derivational morphology learning are relatively rare, contrary to general methods of morphology learning. Roughly, two groups of methods can be distinguished (Walther and Nicolas, 2011): aimed at automated construction of morphological analysers and extraction of morphological models (e.g. segmentation and rules). As we need an analyser of derivatives we focus on the methods of the first group. The vast majority of them is based on unsupervised learning from large annotated corpora. Combi-

nations of different methods of statistical analysis are used in order to identify affixes, stems and word form families, e.g. (Schone and Jurafsky, 2001), *Minimum Description Length* concept is often used in discovering segmentation, e.g. (Kohonen et al., 2009), cf overview in (Walther and Nicolas, 2011). Walther and Nicolas (2011) presented corpus-based extraction of derivational rules. Derivative candidates were filtered on the basis of their frequency in the corpus. From 37.5 million token corpus of French 62,158 derivative candidates were extracted, but only 1,511 new derived French lemmas were identified after ranking candidates. During manual evaluation of a small sample of 100 lemmas: 42 lemmas and relations were identified as correct, but 43 lemmas were definitely incorrect (many due to foreign words and typos). However, we can use a limited but manually annotated set of derivational pairs.

For a limited training data *memory based learning*, e.g. (van den Bosch and Daelemans, 1999), and *transformation-based learning* paradigms, e.g. (Oflazer et al., 2001), were used. The latter was applied to Polish, but without evaluation. Both approaches are valuable options, but Polish derivational rules can be described in terms of prefixes and suffixes added to the derivative base lemma together with a limited set of *internal stem alternations* (Rabiega-Wiśniewska, 2009). Polish derivatives are rarely built by simultaneous use of a prefix and suffix. Transducer based morphological models can easily cope with suffixation and converted to morphological guessers applying recorded rules to unknown words, e.g. for Polish (Daciuk, 2001) and a large scale Polish guesser called *Odgadywacz* (Piasecki and Radziszewski, 2008) of high precision and recall. Transducer guessers have problems with alternations (store exact word parts) and with prefixes (are mostly built on the basis of *a tergo* indexes). In Sec. 4 we present solutions to both problems.

3 Derivative Relations in *plWordNet*

plWordNet is the largest publicly available wordnet of Polish. The version 1.0 was published in 2009 (Piasecki et al., 2009). *plWordNet* 2.0 project started in 2009, according to the contemporary estimates, it is to reach the size of 140–150 thousands of LUs and more than 100000 synsets by the end of 2012. *plWordNet* 2.0 has been extended not only in the number of LUs but also in the number

of lexico-semantic relations. Among them derivationally motivated relations were expanded from two coarse grained defined in *plWordNet* 1.0 to a sophisticated system described in this section.

Word formation is interconnected with semantics: certain senses corresponds to affixes, e.g., English suffix *-er* has several different meanings, among them: ‘(male) agent’ (*thinker, writer, driver, etc.*); ‘instrument’ (*opener, printer, pager*); ‘experiencer’ (*hearer*); ‘stimulus’ (*pleaser, thriller*); ‘patient/theme’ (*fryer, keeper, looker, sinker, loaner*) or ‘location’ (*diner*) (Lieber, 2008, pp. 1-2, 17), cf (Bosch et al., 2008, p. 83). Of course, the same meaning can be carried by many different affixes. The ‘agent’ sense is represented also by *-antl-ent* (*servant, evacuant, descendant*) (Lieber, 2008, pp. 37, 69). The same holds for Polish. Suffix *-ak* is similar in some of its functions to English *-er*, for instance, it has meanings ‘agent’ (*rybak* ‘fisher’, *wieśniak* ‘peasant’, *pływak* ‘swimmer’) or ‘instrument’ (*szczelińiacz* ‘chisel for making cracks’), but it has also few other senses: ‘dweller’ (Polak, Słowak), ‘offspring’ (e.g. *kociak* ‘kitten’, *świnia* ‘piglet’, *kurczak* ‘chicken’), ‘emotional markedness’ (*dzieciak* ‘kid’, *tobuziak* ‘rascal’) (Grzegorczykowa and Pużynina, 1998).

For *plWordNet*, we have chosen relations that have clear semantics and are regular or very frequent in Polish, see (Maziarz et al., 2011a, p. 175), (Maziarz et al., 2011b)). In analysis of the frequencies we followed (Grzegorczykowa and Pużynina, 1998) who based their estimates on the ‘lexicon’ frequencies in (Doroszewski, 1969).

Cross-categorial synonymy is extremely frequent. Semantically they are *transposition* relations causing that the meanings of the words are very close and differ only in their parts of speech.

N-V subtype links deverbal nouns (gerunds) with their bases. We account for regular type on *-anie, -enie, -cie*: *pływanie* ‘swimming’ < *pływać* ‘swim_{impf}’, *napelnienie* ‘filling’ < *napęlnić* ‘fill_{pf}’, *picie* ‘drinking’ < *pić* ‘drink_{impf}’ (Grzegorczykowa and Pużynina, 1998, pp. 393-8).

Pact-V subtype connects active adjectival participle on *-acy* (pact in IPIC tagset) with its verb base (*pijący* ‘drinking’ < *pić* ‘drink’, *pływający* ‘swimming’ < *pływać* ‘swim’). In Polish the active participle may be formed only from imperfective verbs. In many grammars the regular formation is treated as an inflectional form of a given

verb, e.g. (Laskowski, 1998, p. 268), however, we follow Saloni and Świdziński (1998) in considering it an adjective due to its inflectional behaviour.

N-Adj type refers to deadjectival nouns on *-ość*: *bladość* ‘paleness’ < *blady* ‘pale’, *władcość* ‘imperiousness’ < *władczy* ‘imperious’, *małość* ‘smallness’ < *mały* ‘small’; this type is regular.

Markedness (N-N relation) connects related nouns of which one is a *marked* counterpart of the second, unmarked. Three most productive Polish subtypes were included in *plWordNet*.

Diminutives express small size or positive emotional marking and are frequent in Polish. The meaning could be characterised as: ‘*X_{deriv}* is a little or pleasant *Y_{base}*’. The most popular suffixes are *-ek/-ik(-yk)*, *-ko* and *-ka*: *plotek* ‘little or pleasant fence’ < *plot* ‘fence’, *pałacyk* ‘little or pleasant palace’ < *pałac* ‘palace’, *uszko* < *ucho* ‘ear’, *lampka* < *lampa* ‘lamp’ (Grzegorczykowa and Puzynina, 1998, 425-6). Some formations are derivatives from diachronic point of view but we consider only synchronic formations: e.g., word *młotek* ‘hammer’ was derived from *młot* ‘heavy hammer’ with suffix *-ek*, but nowadays it will be ridiculous to say that *młotek* is ‘small or pleasant *młot*’, so this pair is not included in *plWordNet*.

Augmentatives express grand size and negative emotional marking, and may be paraphrased as: ‘*X_{deriv}* is huge or terrible *Y_{base}*’. Augmentative suffixes are, e.g., *-uch*, *-isko(-ysko)* or *-al*: *paluch* ‘huge or terrible finger’ < *palec* ‘finger’, *ptaszysko* ‘huge or terrible bird’ < *ptak* ‘bird’, *nochal* < *nos* ‘nose’ (Grzegorczykowa and Puzynina, 1998).

Young being expresses youth of derivative’s denotat and its paraphrase is: ‘*X_{deriv}* is young *Y_{base}*’. There are two formants: *-e* and *-ak*, e.g.: *małpię* ‘young monkey’ < *matpa* ‘monkey’, *świnia* ‘piglet’ < *świnia* ‘pig’ (Grzegorczykowa and Puzynina, 1998, pp. 429-30).

Femininity (N-N) links nouns denoting women with their male counterparts: *X_{derivate}-Y_{base}* is ‘X is female Y’. For suffixes *-ka* (*pisarka* ‘female writer’ < *pisarz* ‘writer’) the type is almost fully productive, another popular suffixes are *-ini/-yni* (*bogini* ‘goddess’ < *bóg* ‘god’), *-ica* (*kocica* ‘female cat’ < *kot* ‘cat’), *-a* (*markiza* f. ‘marquise’ < *markiz* m. ‘marquis’) (Grzegorczykowa and Puzynina, 1998, p. 422-5). *plWordNet* includes 1745 instances of this relation, see Tab. 1.

Role (N-V) expresses thematic roles of predicate arguments, e.g. agent, object, instrument

etc., see Tab. 1. In the case of the most frequent agent subtype, the most popular suffixes are *-acz* (*spawacz* ‘welder’ < *spawać* ‘weld’), *-ca* (*władca* ‘ruler’ < *właścić* ‘rule’), *-iciel* (*zbawiciel* ‘saviour’ < *zbawić* ‘save’), *-ator* (*restaurator* ‘restorer’ < *restaurować* ‘restore’), there is also (less frequent) backward (paradigmatic) derivation (*szpieg* ‘spy’ < *szpiegować* ‘spy’). Together suffixal and parafigmatic formations account in (Doroszewski, 1969) for not less than 3500 instances (Grzegorczykowa and Puzynina, 1998, pp. 398-416). In *Slowosieć* the relation occurs 4072 times and is most favourite among linguists.

Role inclusion (V-N) in a similar way to *role* refers to thematic roles of predicate arguments which are built into the verb meaning. However verb derivatives include its bases (noun arguments) in *role inclusion*, whereas in *role* noun derivative plays role of argument in its base (predicate, verb). This derivation is relatively frequent in *plWordNet*: 1262 instances. According to (Wróbel, 1998, pp. 577-83) the most frequent subtypes in Polish are *instrument* and *result*, e.g. (*pieprzyć* ‘to pepper’ < *pieprz* ‘pepper’, *dziurkować* ‘to perforate’ < *dziurka* ‘hole’), next *object* (*kartkować* ‘to leaf through’ < *kartka* ‘a sheet’) and *subject* (*sędziować* ‘to referee’ < *sędzia* ‘referee’), the rest are less productive. The assumptions were confirmed by *plWordNet* data statistics, see Tab. 1.

State/feature bearer (N-Adj) and **state/feature** (Adj-N) are both very productive in Polish. The meaning of the relation linking *X_N-Y_{Adj}* could be articulated in following way: X is/has feature Y. The most frequent suffixes (more than 100 lemmas in (Doroszewski, 1969)) are *-ec* (*mędzec* ‘sage’ < *mądry* ‘wise’), *-ka* (*dziczka* ‘rootstock’ < *dziki* ‘wild’), *-ak* (*dziwak* ‘freak’ < *dziwny* ‘strange’), *-ik* (*okrutnik* ‘cruel man’ < *okrutny* ‘cruel’), together with less frequent suffixes these relations are represented by about 600-800 instances (Grzegorczykowa and Puzynina, 1998, p. 420-1). Till now we have introduced 219 feature bearer relations into *plWordNet*.

Inhabitant (N-N) describes X as an ‘inhabitant/dweller of Y’, where Y is the base denotation. Inhabitant names are derived from geographical proper names (for countries, regions, cities, towns, villages and parts of the world) with such frequent suffixes as *-anin* and *-czyk* or with paradigmatic backward derivation: *Afrykanin* ‘African’ < *Afryka* ‘Africa’, *Wietnamczyk* ‘Vietnamese’ <

Relation	Subtype	Instances	Precision [%]	Recall [%]
aspectuality	-	5835	99.0	75.3
derivativity	-	1752	77.4	30.5
feature bearer	-	176	40.0	14.2
femininity	-	1458	96.5	57.4
inhabitant	-	153	71.7	14.3
state	-	188	40.8	9.6
markedness	deminutivity	1286	96.9	60.6
markedness	augmentativity	213	67.9	23.9
markedness	young being	46	26.7	28.5
semantic role	agent of hidden predicate	1374	89.4	30.5
semantic role	agent	625	58.9	22.2
semantic role	time	45	11.7	59.5
semantic role	location	129	24.2	26.4
semantic role	location of hidden predicate	143	59.2	25.8
semantic role	instrument	327	29.7	36.6
semantic role	patient	92	26.7	31.7
semantic role	other	82	0.0	41.2
semantic role	product of hidden predicate	39	35.0	34.2
semantic role	product	614	32.7	54.7
cross-categorial synonymy	N-ADJ	1219	98.5	82.2
cross-categorial synonymy	N-V	1173	82.1	79.1
cross-categorial synonymy	PACT(ADJ)-V	99	95.9	66.9
role inclusion	other	59	15.0	42.0
role inclusion	agent inclusion	108	12.8	20.2
role inclusion	time inclusion	14	10.0	45.0
role inclusion	location inclusion	33	20.0	42.5
role inclusion	instrument inclusion	287	38.8	43.2
role inclusion	patient inclusion	79	10.0	35.0
role inclusion	product inclusion	323	38.8	34.4

Table 1: Fine-grained cross-validation results and number of instances.

Vietnam ‘Vietnam’, *Bulgar* ‘Bulgarian’ < *Bułgaria* ‘Bulgaria’ (Grzegorczykowa and Puzynina, 1998, pp. 437-8) (147 instances in *plWordNet*).

Aspectuality (V-V) expresses aspectual and *Aktionsart* differences. The phenomenon is regular in Polish. Two subtypes are defined in *plWordNet*: 9145 instances of pure aspectuality (only aspectual differences, *wykopać* ‘to dig_{pf} sth up’ - *wykopywać* ‘to dig_{impf} sth up’), 3979 instances of secondary aspectuality (aspect differences + lexical meaning shift, *zaświecić* ‘to start_{pf} shining’ - *świecić* ‘shine_{impf}’).

Derivationality groups all derivational relations that are not included in the above subtypes.

4 Derivator

As regards a number of derivative pairs described in *plWordNet*, we followed supervised learning scheme in the construction of an analyser of derivatives: trained and next used to suggest derivational relation instances during expansion.

4.1 Learning

Polish derivative relations are encoded in a similar way to morphological oppositions: mostly en-

coded by a suffix and only some of them by a prefix. *Odgadywacz* – a morphological ‘guesser’ generates morphological descriptions for unknown Polish word forms with relatively high accuracy on the basis of suffixes learned from annotated examples. Here, we want to expand *Odgadywacz* and its learning model to recognition of derivational relations. *Odgadywacz* was described in (Piasecki and Radziszewski, 2008), below a brief overview is presented. The guesser is based on a deterministic transducer which takes a word reversed sequence of letters and returns morphological description attached to the terminal node. Only an *artificial suffix* is read, that was identified during learning. Training data: ⟨word form, morphosyntactic tags, base form⟩ were acquired from a large annotated corpus of Polish.

The guesser learning is divided into two phases: transducer tree building and pruning. During the first phase for each reversed word form a transducer path is built, partially utilising already existing nodes as long as different suffixes overlap. Morphosyntactic tags and base form reconstruction specification are recorded in the terminal nodes cf (Piasecki and Radziszewski, 2008). Next

the tree is pruned. All final non-branching path parts are cut off, next branches with a small number of training examples attached are further cut off and the information attached is copied.

We used *Odgadywacz* as a core to build upon it a tool, called *Derywator* (a ‘derivator’ – generator of derivatives) which returns possible derivative bases and the relation names for a lemma on the input. Accordingly, a learning triple must now be: ⟨a derivative, a relation tag, a derivative base⟩. Two problems emerged: internal stem alternations and construction of derivatives on the basis of both: suffixes and prefixes. The latter was solved by training two guessers working in parallel in reversed directions, see the algorithm below.

Transformation of a base into a derivative is often not only achieved by suffixation or prefixation but also by up to several internal stem alternations. Stem alternations are not supported by *Odgadywacz* (are less frequent in morphological forms). To model the alternations, we search for a sequence of alternations that make the derivative base overlapping on its ending or beginning with a derivative. The identified alternations are next used to transform the bases before the training data are delivered to *Odgadywacz*. The sequences are linked to the guesser tree nodes and used for reconstructing derivative bases during guessing.

Learning algorithm

Input: $L = \langle \text{a derivative, a relation tag, a derivative base} \rangle$, T – table of alternations (a mapping: a letter sequence to a letter sequence)

For each $e = \langle d, r, b \rangle \in L$:

1. t_p = sequence of at most k substitutions from T such P is shared beginning of d and b .
2. t_s = as in the above but in the relation to a shared ending S
3. If $\text{length}(P) \geq \text{length}(S)$
 - then add $\langle d, r + t_p, t_p(b) \rangle$ to the training examples of the normal guesser,
 - else add $\langle \text{rev}(t_s(b)), r + t_s, \text{rev}(d) \rangle$

In steps 1–2, the longest ending and beginning shared by the derivative and its base are identified. But, we assume that alternations can occur in any position and T (127 possible alternations defined by a linguist) is used to sequentially extend the found shared beginning and ending, e.g. for the derivative *lwiarnia* ‘≈lion enclosure’ and its base *lew* ‘lion’ the mapping [’lə’ / ’l’] is added to

t_p : $b = \text{'lw'}$ and $P = \text{'lw'}$. In step 3 the type of the guesser to be trained is chosen depending whether the beginning (the normal guesser) or the ending is longer. The former is applied in the above example, the latter in the case, e.g., *zrobić* ‘to do_{perf}’ – *robić* ‘to do_{imperf}’. As *Odgadywacz* assumes that the forms delivered on input are differentiated by the suffix only, the prefix-based guesser is simulated by reversing the letter order in both: the derivative and its transformed base.

4.2 Derivative Base Recognition

For a possible derivative both must be guessed: its derivative base and the relation type. There is no information which *Derywator* module to apply: the suffix or the prefix-based. So, both modules are used in parallel and the results are filtered. The filtering is based on *Morfeusz SGJP*¹ – a Polish morphological analyser, as common lemmas are only important for the wordnet expansion. Morphological filtering is based on the observation that for most derivational relations we can find constraints on the acceptable morphological characteristics of the relation instance constituents, e.g. for *femininity* only nouns in the nominative case are accepted on both sides and the derivative must be in the female gender.

Recognition algorithm

Input: a lemma l , *Derywator* modules, R – morpho-syntactic filtering rules.

1. l is delivered to both modules that return a set of triples: $\langle b, t, r \rangle$ where b is a base reconstructed by the guesser algorithm, t a sequence of substitutions associated with the guesser node during learning, r – relation tag.
2. For each triple:
 - (a) b is transformed by the reversed sequence of substitutions t .
 - (b) if b was generated by the prefix-based guesser than it must be reversed.
3. Triples: $\langle l, r, b \rangle$ are filtered:
 - l and b are first morphologically analysed – non recognised are discarded,
 - and next all descriptions (forms can be ambiguous) are compared with the filtering rules for r – at least one pair of description variants must match the rule.

¹<http://sgjp.pl/morfeusz/>

Relation	Precision [%]	Recall [%]
markedness	97.1	55.8
semantic role	75.7	35.6
c-c synonymy	90.5	80.4
role inclusion	100.0	36.4

Table 2: Coarse-grained cross-validation results.

Morphological filtering limits the intrinsic lexical over-generation of the guessers and is always used. Very often non-words or non-lemmas are generated as potential derivative bases, especially for input lemmas that are not derivatives.

Derywator trees combine unique suffixes and prefixes together with substitution sequences of contextually constrained alternations. However, all this is still insufficient to recognise unambiguously the derivational relation: a lemma pair can be not a derivational pair, e.g. *Fryzyjczyk* ‘Frisian’ – *fryz* ‘frieze’, represent a different relation, e.g. *przepychaczka* ‘declogger’ is not a feminine from *przepychacz* ‘plunger’ or are valid only for the specific LUs, e.g. *kometka* 1:‘small comet’ or 2:‘badminton’ – *kometka* ‘comet’. Semantic information must be considered to decide. For some relations we can refer to the general semantic restrictions based on semantic domains or upper level hypernyms, e.g. for *femininity* both: a derivative and the base must belong to the same domain: *os* (persons) or *zw* (animals), or both must be hyponyms of *<istota 3 ‘being’*. For some relations such simple rules does not exist, e.g. a derivative in *semantic role*: *patient* or *object* can be anything. Generated pairs in which only one lemma is in the wordnet are most interesting for wordnet expansion. But, they are only partially covered by the simple semantic filtering. The filtering must be based on semantic data extracted from the corpus.

5 Evaluation

Evaluation was performed on plWordNet from 06.09.2011. cf Tab. 1. First, a modified 10-fold cross validation was performed, see average results in Tab. 1: each subtype was randomly divided into 10 parts, one part from one subtype was used for testing. For many subtypes the results are low, contrary to the coarse grain level of the whole types, cf Tab. 2. Thus, the majority of errors are made on the subtle distinctions between subtypes.

Next, we used a list of 341230 lemmas (N, gerunds, Adj, V) from *Morfeusz SGJP*. in two

Pair type	Basic t-set	Extended t-set	Gain
Instance	10582	11105	523
Derivative	37886	41333	3447
Derivative – monosemous base	24681	26727	2046
Base	5905	6467	562
Both	103990	107695	3705

Table 3: Wordnet expansion learning gain.

experiments. In the first experiment, two *Derywator* versions were trained: on the training data from *plWordNet* the version 18.08.2011 (15718 examples) – *basic training set* – and the version 06.09.2011 (17971 examples, 14% more) – *extended training set*, i.e. the one used in Tab. 1 and 2). Both versions were next applied to the whole lemma list. The results are presented in Tab. 3: gain means the increase in the number of new derivational pairs. Five classes were identified: new *instance* of a derivational relation – both elements already in *plWordNet*, new *derivative* – the base is in *plWordNet*, but not the derivative, new *derivatives with monosemous base* are of special importance for the wordnet expansion as they do not require sense disambiguation of base LU and new *base*. The last class includes pairs with both elements new. The smallest gain (about 5%) was observed for new instances, however, this is a more diagnostic tool – they should be already described. However, for the cost of manual addition of 2253 we obtained 8237 potential new instances. Moreover, *Derywator* should support consistent and systematic description of the derivational relations. The training examples are a standard result of the wordnet editing, no extra work is required.

Among 41333 potential new instances in the new *derivative* class 26727 refer to the monosemous bases (with one synset). Precision for this subclass was checked manually. For each subtype a sample of at most 50 pairs was randomly selected. Linguists were asked to assign the pairs to the three classes: *Correct Subtype*, *Correct Type* (of any subtype of the same type), *Derivatives* (a different type of derivation), *non-derivational* pair, see Tab. 4. Concerning the limited number of the training examples, the precision of the identification of derivatives is on a good level, mostly above 80%. Only for two subtypes, the precision is unacceptable: *inhabitant* – is based on a limited set of suffixes that produce good results only for lemmas representing persons – and *marked-*

ness: *augmentativity* – the dictionary of *Morfeusz* is too broad and includes even tokens unknown to Google. In the case of four subtypes the general precision is lower, e.g. *derivativity* – very varied training examples. Many errors were caused by the over-generation of *Morfeusz*.

The comparison of the general high precision with these of the subtype level shows problems with precise differentiation among different kinds of semantic associations.

6 Toward Wordnet Expansion

Derywator trained on the wordnet enables automated wordnet expansion with derivational relations. The generated new instances show potential gaps in the wordnet (a diagnostic tool), but they cannot be automatically added to the wordnet – lemma pairs produced by *Derywator* must be mapped to LUs and not all of them are proper instances of the signalled relation e.g., for *osad–osadzić* described by *Derywator* as *role:patient* there are several LUs: *osad 1* ‘deposit’, *osadzić 1* ‘to plant’, *osadzić 2* ‘to set’, *osadzić 3* ‘to stop short’ and *osadzić 4* ‘to settle’. However, only *osad 1 –role:patient– osadzić 4* is a correct link. Moreover we can notice that *osad 2* ‘trace’ (metaphorical) is lacking and it cannot be linked to any of the four LUs for *osadzić*. But several LUs can be lacking, e.g., a lemma pair *podbudowa, podbudować* returned as a *role:tool* instance, corresponds to LUs: *podbudowa 1* ‘foundations’ – *podbudować 1* ‘to support’, a correct instance, but also to *podbudować 2* ‘underpin’ for which a noun LU *podbudowa 2* ‘base of something, e.g. a road’ is lacking and must be added. In addition, one more verb LU is absent: *podbudować 3* ‘to reassure somebody’, which is not related by *role:tool* to any of the two LUs of *podbudowa*.

Derywator must be combined with automated assignment of new lemmas to the wordnet structure, cf also (Koeva et al., 2008). We plan to extend the WordnetWeaver system (Piasecki et al., 2009). As the mapping of new lemmas to the wordnet synsets is not sufficient for precise filtering of erroneous relation instances, lemma pairs must be classified according to the represented relation on the basis of knowledge extracted from a large corpus. If full automation is not possible, at least more detailed semantic filtering on the basis of the corpus extracted information can also delimit suggested relation instances and decrease the

error rate to the level acceptable by linguists.

Morphological filtering rules and semantic ones appeared to be simple but very helpful diagnostic tools. Filtering of training data revealed some errors in *plWordNet*. Only a small portion of rejections were derivatives based on simultaneous suffixation and prefixation, e.g., compounds.

7 Further Research

The approach is relatively simple: ready-to-use morphological guesser and analyser, and about one person month for expanding it to a tool learning productive derivational rules of good overall accuracy. The only language dependent element is the list of possible internal stem alternations. However, the generation of derivatives is only half-way to the effective semi-automatic wordnet expansion. A semantic, corpus-based filtering and sense identification techniques are required. Appropriate selection of the new manually described examples in a way maximising the expected gain achieved in training *Derywator* on extended data is needed in order to optimise workload and guarantee training space exploration.

Acknowledgments Work financed by the Polish Ministry of Education and Science, Project N N516 068637.

References

- S. Bosch, Ch. Fellbaum, and K. Pala. 2008. Enhancing wordnets with morphological relations: A case study from czech, english and zulu. In *Proc. of the Fourth Global WordNet Conference. GWC 2008*.
- J. Daciuk. 2001. Computer-assisted enlargement of morphological dictionaries. In *Proc. of Finite State Methods in Natural Language Processing Workshop, 13th ESSLLI, Helsinki August 2001*.
- W. Doroszewski, editor. 1969. *Słownik języka polskiego*, volume I-X. PWN, Warszawa.
- Ch. Fellbaum, A. Osherson, and P.E. Clark. 2009. Putting Semantics into WordNet’s “Morphosemantic” Links. In Z. Vetulani and H. Uszkoreit, editors, *Human Language Technology. Challenges of the Information Society, LTC’2007, Revised Selected Papers*, volume LNCS, pages 350–358. Springer.
- R. Grzegorczykowa and J. Puzynina, 1998. *Gramatyka współczesnego języka polskiego. Morfologia*, volume 2nd, chapter IV: Słowotwórstwo. Rzecznik, pages 389–468. PWN.
- S. Koeva, C. Krstev, and D. Vitas. 2008. Morphosemantic relations in wordnet – a case study for two slavic languages. In A. Tanács, D. Cséndes,

Relation	Subtype	Precision [%]		
		Subtype	Type	Derivatives
derivativity	-	60.0	60.0	64.0
feature bearer	-	34.0	34.0	88.0
feminity	-	74.0	74.0	92.0
inhabitant	-	8.11	8.11	37.8
state	-	46.0	46.0	88.0
markedness	deminutivity	72.0	74.0	84.0
markedness	augmentativity	32.0	36.0	44.0
markedness	young being	10.5	78.9	81.6
semantic role	agent of hidden predicate	40.0	42.0	70.0
semantic role	agent	45.2	61.9	73.8
semantic role	time	20.0	20.0	100.0
semantic role	location	43.2	61.4	79.5
semantic role	location of hidden predicate	40.0	40.0	86.0
semantic role	instrument	42.0	58.0	90.0
semantic role	patient	44.0	64.0	90.0
semantic role	other	10.0	10.0	92.0
semantic role	product of hidden predicate	23.3	23.3	70.0
semantic role	product	12.0	12.0	96.0
cross-categorial synonymy	N-ADJ	81.6	81.6	87.8
role inclusion	other	45.8	60.4	79.2
role inclusion	agent inclusion	22.0	60.0	72.0
role inclusion	time inclusion	0.0	66.7	66.7
role inclusion	location inclusion	0.0	84.2	84.2
role inclusion	instrument inclusion	20.0	84.0	86.0
role inclusion	patient inclusion	8.0	76.0	76.0
role inclusion	product inclusion	40.0	66.0	76.0

Table 4: Manual evaluation of pairs with new derivative and monosemous base.

- V. Vincze, Ch. Fellbaum, and P. Vossen, editors, *Proc. Fourth Global WordNet Conf.*, pages 239–253.
- O. Kohonen, S. Virpioja, and M. Klami. 2009. Al-lomorfessor: towards unsupervised morpheme analysis. In *Evaluating systems for multilingual and multimodal information access, 9th Workshop of the CLEF*, pages 975–982, Berlin, Heidelberg. Springer.
- R. Laskowski. 1998. *Gramatyka współczesnego języka polskiego. Morfologia*, volume 1st, chapter III: Flexja, pages 151–269. Wydawnictwo Naukowe PWN.
- R. Lieber. 2008. *Morphology and Lexical Semantics*. Cambridge University Press.
- M. Maziarz, M. Piasecki, J. Rabiega-Wisniewska, and S. Szpakowicz. 2011a. Semantic relations among nouns in polish WordNet grounded in lexicographic and semantic tradition. *Cognitive Studies Études Cognitives*, 11:161–181.
- M. Maziarz, M. Piasecki, S. Szpakowicz, J. Rabiega-Wisniewska, and B. Hojka. 2011b. Semantic relations between verbs in Polish Wordnet 2.0. *Cognitive Studies*, 11:183–200.
- K. Oflazer, S. Nirenburg, and M. McShane. 2001. Bootstrapping morphological analyzers by combining human elicitation and machine learning. *Computational Linguistics*, 27:59–85.
- K. Pala and D. Hlaváčková. 2007. Derivational Relations in Czech WordNet. In *Proc. Workshop on Balto-Slavonic NLP*, pages 75–81, Prague.
- M. Piasecki and A. Radziszewski. 2008. Morphological prediction for polish by a statistical *A Tergo* index. *Systems Science*, 34(4):7–17.
- M. Piasecki, S. Szpakowicz, and B. Broda. 2009. *A Wordnet from the Ground Up*. Wrocław University of Technology Press. www.site.uottawa.ca/~szpk/pub/A_Wordnet_from_the_Ground_Up.zip.
- J. Rabiega-Wiśniewska. 2009. On the root-based lexicon for polish. In M. Marciniak and A. Mykowiecka, editors, *Aspects of Natural Language Processing*, volume 5070 of *LNCS*, pages 61–82. Springer Berlin / Heidelberg.
- Z. Saloni and M. Świdziński. 1998. *Składnia współczesnego języka polskiego*. Wydawnictwo Naukowe PWN, Warszawa, 4th, changed edition.
- P. Schone and D. Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *NAACL*. ACL.
- A. van den Bosch and W. Daelemans. 1999. Memory-based morphological analysis. In *ACL*.
- G. Walther and L. Nicolas. 2011. Enriching morphological lexica through unsupervised derivational rule acquisition. In *Proc. of the Inter. Workshop on Lexical Resources (WoLeR) at ESSLI. Ljubljana*.
- H. Wróbel, 1998. *Gramatyka współczesnego języka polskiego. Morfologia*, volume 2nd, chapter IV: Słownictwo. Czasownik, pages 389–468. PWN.

Mass noun classifiers in Nepali

Madhav P Pokharel
Tribhuvan University, Nepal
madappokhrel@gmail.com

Abstract

Nepali stands out among the classifier languages of the world, because it is the only language to our knowledge, which uses extensive number of mass noun classifiers. Numeral classifiers of other classifier languages typically supply classifiers only for the count nouns which have a definite boundary, but Nepali has conceptually bounded units as mass noun classifiers, abstract noun classifiers, verbal (action and state) classifiers and also classifiers for pieces in addition, of course, to classifiers for counting naturally bounded units. However, this paper is limited to the mass noun classifiers in Nepali.

1. Introduction

Among the classifier languages of the world, Nepali stands out for using mass noun classifiers in addition to count noun classifiers which are commonly found in all the classifier languages. In a typical classifier language like Burmese ((Latter, 1845) (Bur651) (Pe, 1965) (Becker, 1975)), Thai (Haas, 1942), Chinese ((Schafer, 1948) (T'sou B. K., The place of classifiers in a generative grammar of Chinese, 1965)), Vietnamese (Nguyen, 1957), Khmer (Jacob, 1965), Sino-Tibetan (Hashimoto, 1977) and Japanese (Matsumoto, 1985), there are classifiers to count bounded objects, but in Nepali although there are classifiers to count bounded round and unround objects like man, mammal, insect, similar to other noted classifier languages, there are many mass noun classifiers in addition to these count noun classifiers and the number of those mass noun classifiers exceed the number of count noun classifiers.

2. Classifier

According to Pe (Pe, 1965) the word ‘classifier’ was coined by Latter (Latter, 1845).

‘Classifier’ is an extra free or bound (typically obligatory) morpheme that co-occurs with a noun in a noun phrase and it states the semantic features of the noun it co-occurs with.

Malinowski (Malinowski, 1929) supplies six criteria for defining ‘classificatory particles’ [classifiers]:

- a. Is the numerical classification comprehensive or not?
- b. Must the numerals be used obligatorily with classifiers?
- c. What are the rules of sporadic use?
- d. Does the classification embrace all nouns or only a few isolated groups?
- e. How many classifying formatives do there exist?
- f. Are the examples exhaustive or nearly so, or only a small fraction of the full list?

There are more than 500 classifiers in the data. Almost all the count nouns in the noun phrase obligatorily take a classifier while counting it. Mass nouns typically fall outside this counting (Greenberg, Numeral classifiers and substantival number: problems in the genesis of a linguistic type, 1972). However, mass noun classifier is a device in Nepali to supply units to count culturally functional natural sections of mass nouns.

In addition to making a classificatory device for counting culturally relevant sections of mass nouns, there are classifiers for the conceptually bounded units of abstract nouns and verbal nouns denoting actions and states in the data. However, abstract noun classifiers and action and state verbal noun classifiers (including the most typical count noun classifiers) fall outside the focus of this paper.

According to Greenberg (Greenberg, Studies in numeral systems, I: Double numeral systems, 1974) the classifier morpheme disappears when someone tries to translate a text from a classifier language like Nepali or Japanese to a non-classifier language like English.

- a. Nepali: <ek koso kera> (one-CL.long.object-kera) ‘a banana’
- b. Japanese: <banana ip-pon> (banana-one-CL.long.object) ‘a banana’

The examples show that the classifiers for long object <*koso*> in Nepali and <*hon*> in Japanese are lost when the noun phrase is translated into English from each of the languages.

This stance presumes that classifiers do not supply any meaning to the noun it qualifies, but T'sou (T'sou B. K., The place of classifiers in a generative grammar of Chinese, 1965) and Becker (Becker, 1975) differ from Greenberg (Greenberg, Studies in numeral systems, I: Double numeral systems, 1974). Our data also support T'sou and Becker. (Kay, 1971)

3. Numeral classifier as a subtype of classifiers

There are six types of classifiers found in the languages of the world (Aikhenvald, 2000); (Grinevald, 2000). They are:

- Gender or noun class
- Numerical classifier
- Noun classifiers
- Verb classifiers
- Genitive or possessive classifiers
- Locative or deictic classifiers

4. Three types of classifier in Nepali

Nepali has three types of classifier, viz.

- Gender or noun class
- Verb classifiers and
- Numerical classifiers.

Mass noun classifier is a subtype of numerical classifiers in Nepali.

5. Quantifiers vs. mass noun classifiers

Every language has a set of words called quantifiers. They are also called measure words. A numerical classifier typically occupies the syntactic slot chosen by a quantifier like kilo, meter, liter, etc; therefore a quantifier may be confused to be a classifier. Care must be taken that a quantifier may not be counted in the list of classifiers.

A measure word or a quantifier does not qualify the inherent property of the noun with which a quantifier collocates, e.g.

- Ek kilo pani* 'a kilo of water'
- Ek kilo phalām* 'a kilo of iron'
- Ek kilo kera* 'a kilo of banana'

But a classifier (in our case a typical mass noun classifier) classifies the inherent semantic properties of the noun it qualifies and collocates with, e.g.:

- Ek thopo pani* 'a drop of water' (while the drop is hanging at the source)
- Ek thoplo pani* 'a drop of water' (after the drop drops on a plane surface)
- Ek t^{sh}ito pani* 'a drop of water' (when the drop is slanted upwards or downwards)
- Ek t^{sh}irko pani* 'a drop of water' (when the drop falls on a surface being split up and sprinkled into a cluster of droplets)
- Ek tapkjani pani* 'a drop of water' (when one of the several drops of water periodically falls inside the house leaking through the roof)
- Ek bat^{sh}ito pani* 'a drop of water' (slanted and strayed raindrop)

Each of the classifiers in the examples means 'a drop', however, each of the drops define different orientations of the drop. The classifier <*t^hopo*> is used when the drop is hanging; <*t^hoplo*> is used when the drop falls on the surface; <*t^{sh}ito*> is used when the projectile of the drop of is slanted (both upwards and downwards); <*t^{sh}irko*> is a sprinkled cluster of drops; <*tapkjani*> is every drop of water periodically leaking through the roof or ceiling and <*bat^{sh}ito*> is the raindrop slanted downwards. Each of them occupies a typical classifier slot. We claim that each of the drops in the examples is a classifier not a quantifier.

T'sou ((T'sou B. K., The place of classifiers in a generative grammar of Chinese, 1965) (T'sou B. K., The structure of nominal classifier systems, 1973) (T'sou B. K., 1976)) has supplied two semantic features [ENTITY] and [EXACTNESS] to define classifiers. According to him quantifiers are [-ENTITY,+EXACT], abstract nouns are [-ENTITY,-EXACT], classifiers like Chinese <*zhi*> 'chicken' are [+ENTITY,+EXACT] and group classifiers like '(a) school (of fish)' are [+ENTITY,-EXACT]. Following T'sou Nepali mass noun classifiers fall under the fourth group.

6. Model of analysis

We have followed here the model and methodology of analyzing classifier semantics adopted by Berlin ((Berlin B. &, 1964) (Berlin B. , Tzeltal numeral classifiers. A study in ethnographic semantics, 1968) (Berlin & Breedlove, General principles of classification and nomenclature in folk biology, 1973)), but the method of classificatory device is somewhat influenced by Simpson's (Simpson, 1945) biological classification and Kay (Kay, 1971).

7. Description of mass noun classifiers

Following is the classified list of mass noun classifiers in Nepali. Each of the classifiers occupies the syntactic slot typically occupied by a classifier/quantifier. Besides, each of such mass noun classifiers classifies the fine grained properties of the noun it qualifies.

1. Granules/ powder/ flour

1.1. Ashes or sand etc. for rubbing the pots clean
 <masko>

1.2. Salt

1.2.1. Big crystal <dikko>

1.2.2. A pinch of salt <tsimt-i> '(literally) forceps'
 (metaphorical)

2. **Semisolid** [Characteristic test against both solid & liquid: <tsoli>] (prototype <bat> 'rice' / <numi> 'butter' / <mat> 'mud')

2.1. Edge prominent slice or chop

2.1.1. Small thickness <tsoil-o>/ <tsoil-i>

2.1.2. Bigger chop of unmarked thickness <tsappari>

2.2. Unmarked for edge

2.2.1. Small lump <qallo>

2.2.2. Bigger lump <bakkano>

2.2.3. Still bigger and harder mass of clay <jisko>

2.3. Particular place where people go to dig typically red color mud to paint their house wall, mud floor and inside and outside of the door
 <maqkeno>

2.4. Stool [these are different classifiers for stool of mammals/ category is stool]

2.4.1. Buffalo <thas>,

2.4.2. Cow <thapro>,

2.4.3. Pig, cat, tiger, mouse <lid>, (semi cylindrical shape)

2.4.4. Horse <lidi>

2.4.5. Goat <bakulo>

3. **Emulsion, suspension and paste** [Characteristic test against semisolid and liquid: <lakto> / <pitko> (vs. <turko> for liquid) 'smallest natural drop/ mass/ quantity'

3.1. Edge prominent <dikko> (feels solid)(e.g. <dahi> 'yoghurt'; <khakar> 'phlegm')

3.2. Dispersed and without edge <lakto> (does not feel like solid) (<sinjan> 'nasal mucous', <dahi> 'yoghurt', <atsar> 'pickle')

3.3. Length prominent and smaller than referred to by <lakto> (e.g. toothpaste) <litko>

3.4. Quantity of ghee that can be scooped up with an index finger hook <aulo> 'finger'

3.5. Not dispersed smallest natural mass (without edge) (usually collocates with food items) <pitko>

3.6. Solidified chunks of blood <palso>

3.7. Chunks of solid particles separated/ not separated from liquid (related to food or drink) <shokro> (maybe etymologically related with <tsokto> 'piece of meat or forcibly torn piece of cloth)

4. **Liquid** [Characteristic test against emulsion: <turko>] <pani>, <dud>, <tel>; <turk-o>/ <turk-i> vs. <pitko>

4.1. Static (source & container classifiers) <kolo> 'stream', <kulo> 'canal', <bangalo> 'estuary', <dha> 'lake', <qbilkko>/ <khobilto> 'pit', <mul> 'source'

4.2. Quantity of water occupied in two palms joining together <dzuli>

4.3. Dynamic (direction: upward, downward, horizontal, slanted

4.3.1. Direction:

4.3.1.1. Upward <bulk> 'boil'

4.3.1.2. Downward

4.3.1.2.1. Drop

4.3.1.2.1.1. Divergent single <tsbit-o> 'sprinkle' (?)

4.3.1.2.1.2. Sprinkled drops (multiple) drops or spots made by them <tsbirko> (etymology * <tsbit-k-o> → <tsbir-k-o>) (<k> 'bounded instantiation of an action or event)

4.3.1.2.1.3. Convergent

4.3.1.2.1.3.1. Hanging on the source <thopo>

4.3.1.2.1.3.2. Fallen on some object <thoplo>

4.3.1.2.2. Rain shower

4.3.1.2.2.1. Big shower <dakko>

4.3.1.2.2.2. One shower (normal) <dzar>

4.3.1.2.2.3. Tap/ line of liquid while pouring <daro>

4.3.1.3. With force

4.3.1.3.1. Sudden surge

- 4.3.1.3.1.1. Usually dirty water/ smoke <muslo> '(literally) pole of flour mill for crushing grains' (metaphorical)

- 4.3.1.3.1.2. Usually clean water <mulko>

4.3.1.3.2. Linear single spray of liquid <sirko>

- 4.4. Number of times a cow or a buffalo gives milk
<sâdz> 'evening'

- 4.5. Quantity of oil needed for one time curry
<buṭun>

5. Gas

5.1. Eye sensitive

- 5.1.1. Pungent smoke around felt in the eyes <p̥atyol>

- 5.1.2. Smoke <kūḍullo> '(literally) coil'

- 5.1.3. Smoke/ dust/ muddy water (with force)
<muslo> '(literally) pole of flour mill for crushing grains' (metaphorical)

- 5.2. Nose sensitive (smell) <h̥rak>

- 5.3. Ear sensitive

- 5.3.1. Whistling in the wind <susel-o> vs. <susel-i>
'whistling' (metaphorical)

- 5.3.2. Sound of wind/ hurricane <huñko>

5.4. Skin sensitive

- 5.4.1. Cold <sireṭo>,

- 5.4.2. Heat <āts>

6. Fire

- 6.1. Spark <đzilko>

- 6.2. Glow <pilko>,

- 6.3. Burning fire particle <p̥ilungo>

- 6.4. Flame <đzwalo>

- 6.5. Biggest flame <dənko>

- 6.6. Burning coal <koilo>

- 6.7. Torch made of braided cloth <pult̥o>,

- 6.8. Tongs <tsimta>

- 6.9. Torch made on the stick <rāko>,

- 6.10. Collection of burning coals <bujro>,

- 6.11. Burning fire stick on one side <agullto>,

- 6.12. Hearth <Ageno>,

- 6.13. Bonfire <duni>

7. Light

- 7.1. (a) Sunlight and (b) dress measured in terms of human body <aj> 'body'

- 7.2. Glow from a smallest light source <pilko>,

- 7.3. Sunrise/ moonrise <đzulko>,

- 7.4. Spark <đzilko> [cf. <đzalko> (for seeing) & <đzadžalko> (for memory)]

8. Conclusions

These examples are sufficient to claim that there are mass noun classifiers and that the mass noun classifier is a separate class and category to be explored in other languages and Universal Grammar.

Bibliography

Aikhenvald, A. Y. (2000). *Classifiers: A typology of noun categorization devices*. New York: Oxford University Press.

Becker, A. L. (1975). A linguistic image of nature: The Burmese numeral classifier system. *Linguistics*, 165, pp. 109–121.

Berlin, B. &. (1964). Descriptive semantics of Tzeltal numeral classifiers. *American anthropologist, new series*, 66.3 (2), 79-98.

Berlin, B. (1968). Tzeltal numeral classifiers. A study in ethnographic semantics. *Janua linguarum, Series 70.* .

Berlin, B., & Breedlove, D. E. (1973). General principles of classification and nomenclature in folk biology. *American anthropologist*, 214-242.

Burling, R. (1965). How to choose a Burmese numeral classifiers. In M. E. Spiro (Ed.), *Context and meaning in cultural anthropology* (pp. 243-264). New York: The Free Press.

Greenberg, J. A. (1972). Numeral classifiers and substantival number: problems in the genesis of a linguistic type. *Working papers on language universals*, 9, 1-39.

- Greenberg, J. A. (1974). Studies in numeral systems, I: Double numeral systems. *Working papers on language universals*, 14, 75-89.
- Grinevald, C. (2000). A morphosyntactic typology of classifiers. In G. Senft (Ed.). New York: Cambridge University Press.
- Haas, M. R. (1942). The use of numeral classifiers in Thai. *Language*, 18 (3), 201-205.
- Hashimoto, M. J. (1977). The genealogy and the role of the classifier in Sino-Tibetan. *Computational analysis of Asian and African languages*, 7, pp. 69-78.
- Jacob, J. M. (1965). Notes on the numerical and numeral coefficients in Old, Middle and Modern Khmer. *Lingua*, 15, pp. 143-162.
- Kay, P. (1971). Taxonomy and semantic contrast. *Language*, 47 (1), 866-887.
- Latter, T. (1845). *A grammar of the language of Burmah*. Calcutta: Thacker.
- Malinowski, B. (1929). Classificatory particles in the language of Kiriwina. *Bulletin of the School of Oriental Studies*, 1 (4), 33-78.
- Matsumoto, Y. (1985). *Japanese numeral classifiers: their structures and acquisition*. MA dissertation, Sophia University, Tokyo.
- Nguyen, D. H. (1957). Classifiers in Vietnamese. *Word*, 13, 124-152.
- Pe, H. (1965). A re-examination of Burmese 'classifiers'. *Lingua*, 15, 163-185.
- Schafer, E. J. (1948). Noun classifiers in Classical Chinese. *Language*, 24 (3), 408-413.
- Simpson, G. G. (1945). The principles of classification and a classification of mammals. *Bulletin of the American Museum of Natural History*, 85, i-xvi, 1-350.
- T'sou, B. K. (1965). The place of classifiers in a generative grammar of Chinese. *Quarterly progress report, MIT: Research laboratory of electronics*, 78, pp. 221-226.
- T'sou, B. K. (1973, January 2-6). The structure of nominal classifier systems.
- T'sou, B. K. (1976). The structure of numeral classifier systems. (& L. P. Jenner, Ed.) *Oceanic linguistics, Special publications* 13 (2), 1215-47.

Acknowledgments

This paper is a part of the research entitled 'Cognitive similarities in the conceptualization of classifiers in Nepali and Japanese' (2009-10) that was conducted in Kobe University Japan under the Japan Foundation fellowship.

Finding a Location for a New Word in WordNet

Paula Pääkkö

University of Helsinki

Helsinki, Finland

paula.paakkoo@helsinki.fi

Krister Lindén

University of Helsinki

Helsinki, Finland

krister.linden@helsinki.fi

Abstract

FinnWordNet is a Finnish wordnet which complies with the structure of the Princeton WordNet (Fellbaum, 1998). It was created by translating all the words in Princeton WordNet. It is open source and contains over 117 000 synsets. We are now testing different methods in order to improve and expand the content of FinnWordNet.

Since wordnets are structured ontologies, a location for a word in FinnWordNet can be pinpointed by its relations to other words. To us, finding a location for a word therefore means finding a hyperonym, a hyponym or a synonym for the word. This article describes some methods for finding a location for a new word in FinnWordNet. Our methods include searching for multiword terms, compounds and lexico-syntactic patterns. Testing shows that with a few simple methods, we were able to find an indicator of the location for 83.2 % of new words. Out of the new synonym pairs we tested, we were able to find an indication for 86.7 %.

1 Introduction

WordNet (Miller et al., 1990; Fellbaum, 1998) is a lexical database for English where words (adjectives, nouns, verbs, adverbs) are grouped into synonym sets, also called synsets. Each synset represents a concept. In addition to synonymy, WordNet also includes other types of semantic relations, for instance antonymy, hyponymy, meronymy and troponymy. The hyperonym/hyponym relation creates a hierarchical structure for nouns. Typically, wordnets are monolingual but lately multilingual wordnets have been under way, see

(Vossen, 1998; Tufis et al., 2004; Pianta et al., 2002)

The Finnish WordNet (FinnWordNet or FiWN) is a Finnish version of the Princeton WordNet (Carlson and Lindén, 2010). There are three main approaches to creating wordnets (Carlson and Lindén, 2010): you can build a wordnet from scratch, translate an existing wordnet or translate the top ontology and extend it with local synonym dictionaries. FiWN was created by translating the word senses in Princeton WordNet 3.0 by using professional translators. The translation process was controlled with regard to quality, coverage, cost and speed of translation. FiWN with translations can also be used as a Finnish–English dictionary.

After translation, our next aim has been to improve and to expand the content of FiWN. One important part of this is to check that FiWN contains the most frequently used Finnish terms and concepts. The downside of creating a wordnet by translation is that the content tends to include terms specific to the source language, in this case terms related to English-speaking cultures, while some central concepts in the target language are possibly left missing.

Another goal is to make sure that the semantic relations as well as the translations are correct. The FiWN search interface¹ also includes a feedback and rating possibility. Crowdsourcing is one of the methods we are using to improve and expand FiWN (Lindén et al., 2012). One important method for enriching FiWN is finding new instances of semantic relations and words in corpora. Lindén et al. (2012) have already established some useful lexico-syntactic patterns for Finnish for finding instances of semantic relations, especially synonyms and hyponyms/hyperonyms.

The next stage of expanding the content of

¹<http://www.ling.helsinki.fi/cgi-bin/fiwn/search>

FiWN is to add new words and relations to FiWN. To us, finding a location for a word in FiWN means finding hyponyms, hyperonyms or synonyms for that word. Hyponyms, hyperonyms and synonyms act as indicators for the location. In this article, we test a few simple methods for finding such indicators. We use synonym pairs found by Lindén et al. (2012) as our test set.

WordNet (Miller et al., 1990) describes hyponymy as follows: A concept represented by the synset $\{x, x', \dots\}$ is said to be a hyponym of the concept represented by the synset $\{y, y', \dots\}$ if native speakers of English accept sentences constructed from such frames as *An x is a (kind of) y*. Inversely, synset $\{y, y', \dots\}$ is the hyperonym of synset $\{x, x', \dots\}$. Miller et al. (1990) also suggest the following definition for synonymy: two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value.

The article is divided as follows. Section 2 describes the related work for finding instances of semantic relations for English. Section 3 describes our methods in more detail. Section 4 describes the material and Section 5 outlines the evaluation and test results, whereas Section 6 discusses the results and future work. Finally, Section 7 draws the conclusions.

2 Related Work

There has been much research for finding word pairs which share a certain semantic relation. These approaches can be divided into two main categories: pattern-based and cluster-based approaches. The latter uses statistics and clustering algorithms to cluster similar words according to context. This approach is based on the *Distributional hypothesis* (Harris, 1968) which states that words that occur in a similar context tend to be similar in meaning. Works using this approach include for instance (Caraballo, 1999), (Lin, 1998) and (Pantel and Lin, 2002).

There are studies with this approach for Finnish as well. In his Ph.D. thesis Piitulainen (2011) did a case study on the distributional similarity of words. He studied nearly twenty thousand nouns that occur often in a Finnish newspaper corpus. Lindén and Piitulainen (2004) introduced similarity recalculation after context clustering to find subclusters and they used translation dictionaries for evaluating the rate of synonymy found in word

clusters.

Pattern-based approaches use lexico-syntactic patterns in order to find the context of a relation. Hearst (1992; 1998) conducted one of the first studies that used patterns for hyponymy relation. Hearst's patterns included for example:

- NP such as $\{NP, NP, \dots, (and|or)\} NP$
"The bowlute, such as the Bambara ndang"
hyponym("Bambara ndang", "bowlute")
- $NP \{, NP\}^* \{, \} or other NP$
"Bruises, wounds, broken bones or other injuries"
hyponym("bruise", "injury"),
hyponym("wound", "injury"),
hyponym("broken bone", "injury")

These manually created patterns usually have good precision but low recall (Hearst, 1998). Hearst also introduced her Lexico-Syntactic Pattern Extraction method (LSPE) which many later methods are based on. She also tried this method for meronymy, but the results were not as promising. The patterns for meronymy were not unique enough but also found other instances of semantic relations. Berland and Charniak (1999) on the other hand got better results since they used more refined statistical measures for ranking the output.

Many automatic and semi-automatic approaches to finding patterns have later been proposed. Many of them are based on Hearst (1992). They usually have the same idea: First gather seed instances of the desired relation and find those occurrences in text. From this, the context determines a new pattern which is then used to find new instances. New instances in turn can be used to find new patterns. Methods differ in how new patterns are evaluated and selected.

Instead of manually created patterns, some methods use machine learning to learn new patterns. These approaches include for example (Snow et al., 2005). They used a supervised learning algorithm to create a hyperonym classifier. Other works are (Girju et al., 2003), (Girju et al., 2006) for finding instances of part-whole relation (meronymy/holonymy) and (Agichtein and Gravano, 2000) for instances of any kind of semantic relation.

Different approaches also use general patterns. These broad coverage noisy patterns have high recall but low precision. Works include (Girju et al.,

2003), (Girju et al., 2006) and (Pantel and Pennacchiotti, 2006). These patterns produce both right and wrong instances and the methods need to filter out the wrong ones. Many methods use statistical evaluation to estimate the accuracy of generic patterns. Works include for example (Brin, 1999), (Agichtein and Gravano, 2000), (Agichtein et al., 2001), (Ravichandran and Hovy, 2002) and (Pantel et al., 2004)

Statistics-based methods are mainly applicable to medium- or high-frequency words, whereas pattern-based methods are applicable also to low-frequency words, because even one occurrence of a pattern is often sufficient. Since most of the remaining words not in FiWN are low-frequency, we will focus on pattern-based methods.

Many have compared their results with wordnet, see e.g. (Hearst, 1992; Lin and Pantel, 2002; Snow et al., 2005). This can be seen as the first step to add words to wordnet because it requires finding a location for an existing word. However, they have not directly addressed the issue of where to add the words that were not in wordnet.

3 Method

The purpose of our study is to find an indication of the location in FiWN for a new word. The indications are hyponyms, hyperonyms and of course, synonyms. Our test set includes synonym pairs for which we try to find a place in FiWN. The synonyms in our material may be single words, compounds or multiword terms. The methods focus on finding locations for nouns or noun phrases.

We have chosen to test and evaluate the following simple methods:

(A) Categorize the word pairs into five groups:

- 1) Pairs where both words are in FiWN and:
 - (a) in the same synset
 - (b) not in the same synset
- 2) Pairs with one word already in FiWN and:
 - (a) the other word can be added to one of the synsets
 - (b) the other word cannot be added to one of the synsets
- 3) Pairs where neither of the words are in FiWN

(B) Check compound words.

(C) Check main word of a multiword term.

(D) Use patterns.

In method A, the synonym pairs can be mechanically compared with the synsets in FiWN and categorized into groups 1, 2 or 3 accordingly. Separating group 2 into groups 2a and 2b needs to be judged by a human.

Group 1a is unproblematic. Group 1b indicates that some meanings are not yet covered in the existing FiWN. As both words in group 1b are already in FiWN, finding a location for the words might seem unnecessary. But now we are actually finding a location for a new meaning of a word. For this group, we can add some assumptions, because we are most likely finding a less frequent or more specialized meaning of one of the words. We assume that either word of the word pair can be added to a synset of the other word or the word can be added as a hyponym of the other word.

Group 2a is straightforward. The synonym pair's meaning is represented by one of the synsets which contains the known word of the word pair. Since different synsets represent different concepts (ergo meanings) it is likely that the new word is added to only one of the synsets. This is reinforced by the fact that most of the high-frequency words are already included in FiWN. This means that if a word is missing from FiWN, it is most likely a rare word, which usually has only one specific meaning.

Group 2b is similar to 1b, but group 2b also implies that the synonym pair's meaning is somehow new. We have two views on this. First, we can assume that the meaning is more precise so that the new word should be added as a hyponym. On the other hand, it is also possible that we have found a new meaning for a word that is already in FiWN.

Method A is used on word pairs whereas methods B-D are used on words. Since our test set consists of synonyms pairs, it is enough that we get an indicator for at least one of the words in a word pair. The other word gets its hyperonym or hyponym via its synonym.

Method B and C rely on a language's innate mechanism of coining new words and terms based on established ones. In Finnish, compound words are written together and the last word is always the main word. This differs from multiword terms, because the main word is not always the last word. For example a *White-tailed Tropicbird* is a kind of bird and a *role-playing game* is a kind of game.

Finding noun compound words is relatively easy in Finnish based on the compound word border. Additionally, proper nouns are excluded from the results to make the results more useful. This leaves out compound person names, for example the surname *Tois#kallio* (*neighbor-hill*). Compound names may reflect concepts, but they are no longer perceived and used that way. In our annotated corpus a hash sign (#) indicates a word boundary in compound words. In addition, we also check if the word includes a dash (-), which is sometimes used as an explicit compound word delimiter, for instance *MIRV-ohjus* (*MIRV missile*).

For the multiword terms we took similar steps as for the compound words, e.g. proper nouns were discarded to exclude person names. Our program chooses the head word of the NP as the hyperonym. If no head word is annotated, then the last noun of the NP is chosen.

Method D uses hand-made patterns. Lindén et al. (2012) established some useful patterns for Finnish which we are using to find a hyperonym or a hyponym for a word in a corpus. These patterns are based on Hearst's patterns (1992; 1998).

4 Material

Lindén et al. (2012) evaluated a few lexicosyntactic patterns. From their subsection of Finnish Wikipedia² articles they found 1405 occurrences of pattern *eli* (*a.k.a.*). They manually checked 1100 and evaluated 583 (53.3 %) as useful (that is to say the occurrence of the pattern produced a word pair with a known semantic relation). Relations were categorized as synonyms, translations (which can be seen as kind a of synonymy) and glosses. From this, we got a test set of 459 synonym pairs.

Since the purpose of these tests is to find a hyponym or hyperonym for the words, we only look at noun pairs. This gives us a test set of 422 unique synonym pairs and 594 unique new words (strings)³.

²<http://fi.wikipedia.org/wiki/Wikipedia:Etusivu>

³The word for a special species of dolphin, i.e. *inia*, was left out as it was incorrectly given the baseform *in* and would have affected the search results of method D by retrieving English text fragments.

4.1 Preprocessing

The corpus we used for finding patterns was Finnish Wikipedia⁴. This corpus was cleaned from the Wikipedia tagging in order to get only the text. The size of the cleaned corpus was 379.4 MB. This corpus was then annotated with Connexor's⁵ fi-fdg -tool⁶. Below is a sample sentence of *Times are hard* in Finnish:

```
1 Ajat aika    subj>2 @NH N PL NOM (Times)
2 ovat olla   main>0 @MAIN V ACT IND
PRES PL P3 (are)
3 kovia kova comp>2 @NH A PL PTV (hard)
4 .
5 <s><s>
```

Our test words were also annotated with the fi-fdg -tool. The annotated words have one word per line with the information for each word in tsv:

1. Word number
2. Surface form of the word
3. Baseform of the word, word boundaries marked with a hash sign (#)
4. Role of the word, e.g. main word or attribute etc.
5. Other annotations, e.g. class, case etc.

Words get their baseform where word boundaries of compound words are also marked. In addition, the main word is marked in multiword terms. These are needed in the methods we are testing.

An example annotation of *urethritis* i.e. *inflammation of the urethra*:

```
1 virtsaputken virtsa#putki attr>2
@PREMOD N SG GEN      (of the urethra)
2 tulehdus tulehdus main>0
@NH N SG NOM           (inflammation)
3 <p><p>
```

5 Results

Results for our methods can be seen in Table 1 and 2.

Group 1 was divided into two groups using FiWN. As we assumed, one word of each word pair in group 1b fit in the same synset as the other word, but not necessarily the otherway around.

⁴<http://dumps.wikimedia.org/fiwiki/>, downloaded January 2011

⁵<http://www.connexor.com/nplib/>

⁶<http://www.csc.fi/english/research/software/fi-fdg>

Group	1a	1b	2a	2b	3	Altogether
Word Pairs	46	14	88	31	243	422
Percentage	10.9 %	3.3 %	20.8 %	7.3 %	57.6 %	100.0 %

Table 1: Results for method A (counting word pairs).

Method	Words	Applicable	Useful	Accuracy
Method B	479	151	119	78.8 %
Method C	479	116	82	70.7 %
Method D	479	4250	601	14.1 %

Table 2: Results for methods B, C and D (counting words).

Adding one word to the other words’s synset required knowing the meaning of the words in order to decide which should be added to which synset, so this was done by hand.

Group 2 was manually checked and divided into two groups. Group 2a were clear cases where the unknown word could be added to one of the synsets of the other word. Some of these words were for instance Finnish terms for international words. For example *kudosoppi* can be added to synset {*histologia*} (*histology*).

Group 2b implied that the word already in FiWN has some new meaning which is why the new word cannot simply be added to an existing synset. This is the case for example in a synonym pair *aurinkokoiras/hevonens* (*sun dog/horse*). Since *sun dog* is a native american term for *horse*, it is more precise and should be added as a hyponym. In group 2b, most of the new words were best suited as hyponyms of the known words.

Only four words from groups 1b and 2b were not suitable as a hyponym or in the same synset of the word in FiWN. For these words we instead used methods B, C and D and found indicators for all of them.

Group 3 is the largest group, which produces most of the new words for FiWN. This means that there is still room for expanding FiWN. On the other hand, at this point we did not exclude proper names from the test words. Test words might include words we are not currently interested in adding to FiWN. These include for example various person names and sport associations.

Excluding pairs from groups 1 and 2, we are left with 243 pairs from group 3, where neither word is in FiWN. The pairs contain 479 unique words. On these words, we used methods B, C and D.

Among the unique words from group 3 our program found 151 compounds. Of these, 119 were

evaluated as having a good hyperonym which gives an accuracy of 78.8 %. There were some clear cases, e.g. *begonia#kasvi* (*begonia plant*) which is a plant. Some were a mere interpretation, for example a legal term *oikeus#olettama* (*legal presumption*) is a certain type of *olettama* (*presumption*).

In group 3 our program found 116 multiword terms and 82 were evaluated as having a good hyperonym. This gives an accuracy of 70.7 %. Words include for example *neoklassinen musiikki* (*neoclassical music*) which is a kind of *musiikki* (*music*). Out of the good hyperonyms, 19 were compound words from the previous method.

The last method was to use patterns from (Lindén et al., 2012). Patterns included:

- *kuten, kuten esimerkiksi* (as/like/such as, as for example)
- *ja/tai muu* (and/or...other)
- *NP(nom)...on/ovat/oli/olivat..NP*
(NP(nominative form)...is/are/was/were...NP)

First we searched for sentences which contained the specific pattern. With the first pattern we found 42360 sentences, with pattern *and...other* 33482 sentences and with pattern *or...other* 5171 sentences. Sentences having the last pattern were 964911.

From these sentences we searched those which contained the test words from group 3. The first pattern matched 340 sentences. With pattern *and...other* we found 266 sentences and with pattern *or...other* 26 sentences. With the last pattern we added a restriction that the word should be in nominative form. This resulted in 3618 sentences.

We manually evaluated which ones indicated a hyperonym or a hyponym. Most of the good results came from the last pattern. Useful patterns in Table 2 indicate how many of the results produced a hyponym or hyperonym for the given word. Testing for compounds and multiword terms resulted in 297 words with no indication of a hyperonym or hyponym. With patterns, we were able to find indications for 142 of those words. We also found new hyperonyms or hyponyms for some of the compound words and multiword terms.

Methods A-D result in only 155 words with no indication of a location in FiWN. On the other hand, since our test set consisted of synonym pairs, we only need an indication for one of the words. This lowered the words with no indication to 100 (50 synonym pairs). Since we had 594 new words (strings), this means that for 83.2 % of all of the new words we found some indication. Out of the 376 new synonym pairs, i.e. excluding group 1a already in FiWN from the total, we were able to find an indication for 86.7 % of the pairs that were new to FiWN.

6 Discussion and Future Work

The tests gave us some insight into which methods are useful in order to find hyperonyms and hyponyms for a given word. Interestingly, finding compound words and multiword terms is simple and the accuracy is far better than using patterns. On the other hand, we currently only tested that a sentence contained both the pattern and the given word. To get better accuracy for patterns, we need to more accurately check that the pattern itself contains the given word.

Categorizing synonym pairs into different groups allows us to concentrate on specific synonym pairs. With a small amount of manual work, we were able to cover groups 1 and 2 and focus solely on group 3 with our remaining methods.

Using both methods B and C for the words allows us to create multilevel hierarchies. For example, the word *laskennallinen virtaus#mekaniikka* (*Computational Fluid Dynamics*) is a multiword term giving us a hyperonym *virtaus#mekaniikka* (*Fluid Dynamics*). In addition, this main word is also a compound word, giving us hyperonym *mekaniikka*⁷ (*Dynamics*).

One problem was the quality of the fi-fdg -tool. Some words were incorrectly annotated, for ex-

ample the wrong word was annotated as the main word. That is why some compounds and multiword terms from methods B and C did not show up in the results.

In group 1b of our synonym pairs, both words were included in FiWN, but they were not in the same synset. We concluded that one of the words in the word pair can be added to the synset of the other word. Deciding on which synset the other word should be added into was done by hand. Later on, this should be automated. For example, if we know which article produced the synonym pair *moira/kohtalotar* (*Moirai, Moirae/Norn, weird sister*) was found in an article talking about Greek gods. *Moirai/Mairae* has a hyperonym *Greek deity* whereas *Norn/weird sister* has a hyperonym *Norn deity*.

Groups 1b and 2b produced new meanings for words already in FiWN. Since those word pairs cover only 10.6 % of our test pairs, we can infer that most of the meanings of the words in FiWN are covered in the current version. We also conclude that categorizing word pairs is a useful method to discover new meanings.

The obvious thing to point out about these methods is that they can be iterated. Even if the hyperonym or hyponym for a word is missing from FiWN, we can use these same methods to find a location for the new hyperonym/hyponym. We can also use the pattern *a.k.a* to find synonyms for the new word.

7 Conclusions

We have described some simple methods for finding a location for a new word in FiWN. This means finding a synonym, a hyponym or a hyperonym for a new word. These indicate where a word can be added. As a side-effect of our first method, we can also find new meanings for some of the known words in FiWN.

Testing showed that for a test set of 594 totally new words (strings) for FiWN, we were able to find an indication of a location for 83.2 % of the words. Generally, exploiting the usual way a language coins new words based on established ones, e.g. compounding and multiword terms, has a better accuracy than general lexico-syntactic patterns. All in all, we were able to find an indication for

⁷Literal translation is *mechanics*.

86.7 % of the pairs that were new to FiWN.

Acknowledgements

We thank the FINCLARIN and METANORD projects for funding our research. We also thank the FinnWordNet participants and the anonymous reviewers for useful comments on the manuscript.

References

- Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries*, pages 85–94.
- Eugene Agichtein, Luis Gravano, Jeff Pavel, Viktoriya Sokolova, and Aleksandr Voskoboinik. 2001. Snowball: a prototype system for extracting relations from large text collections. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, SIGMOD ’01, pages 612–, New York, NY, USA. ACM.
- Matthew Berland and Eugene Charniak. 1999. Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 57–64, Morristown, NJ, USA. Association for Computational Linguistics.
- Sergey Brin. 1999. Extracting patterns and relations from the world wide web. In *Selected papers from the International Workshop on The World Wide Web and Databases*, pages 172–183, London, UK. Springer-Verlag.
- Sharon A. Caraballo. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 120–126, Morristown, NJ, USA. Association for Computational Linguistics.
- Lauri Carlson and Krister Lindén. 2010. FinnWordNet - WordNet på finska via översättning. In *LexicoNordica - Nordic Journal of Lexicography 2010, vol 17, ISSN 0805-2735*.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Roxana Girju, Adriana Badulescu, and Dan Moldovan. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL ’03, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.
- Roxana Girju, Adriana Badulescu, and Dan Moldovan. 2006. Automatic discovery of part-whole relations. *Computational Linguistics*, 32:83–135, March.
- Zellig S. Harris. 1968. *Mathematical Structures of Language*. Wiley, New York, NY, USA.
- Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA. Association for Computational Linguistics.
- Marti A. Hearst. 1998. Automated discovery of WordNet relations. In C. Fellbaum, *WordNet: An Electronic Lexical Database*, pages 131–153. MIT Press.
- Dekang Lin and Patrick Pantel. 2002. Concept discovery from text. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING ’02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL-36, pages 768–774, Morristown, NJ, USA. Association for Computational Linguistics.
- Krister Lindén and Jussi Piitulainen. 2004. Discovering synonyms and other related words. In *Proceedings of CompuTerm 2004, 3rd International Workshop on Computational Terminology*, Geneva, Switzerland, August. Coling.
- Krister Lindén, Mirka Hyvärinen, Jyrki Niemi, and Paula Pääkkö. 2012. Translating WordNet – and then what? *Language Resources and Evaluation*. Submitted.
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.
- Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD ’02, pages 613–619, New York, NY, USA. ACM.
- Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 113–120, Morristown, NJ, USA. Association for Computational Linguistics.

Patrick Pantel, Deepak Ravichandran, and Eduard Hovy. 2004. Towards terascale knowledge acquisition. In *Proceedings of the 20th international conference on Computational Linguistics (COLING-04)*, pages 771–777, Geneva, Switzerland. Association for Computational Linguistics.

Emanuele Pianta, Luisa Bentivogli, and Christian Giardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January.

Jussi Piitulainen. 2011. *Explorations in the distributional and semantic similarity of words*. Ph.D. thesis, University of Helsinki.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 41–47, Morristown, NJ, USA. Association for Computational Linguistics.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1297–1304. MIT Press, Cambridge, MA.

D. Tufis, D. Cristea, and S. Stamou. 2004. Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information Science and Technology, Vol. 7, Nos. 1-2*, pages 9–43.

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Introducing WORDNET in Interpreting Studies

Implications and Desiderata

Francesca Quattri, Prof. Chu-Ren Huang

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University (PolyU)

Abstract

Multi-word units (MWUs), interpreted in these studies as both collocations and phrasemes, pose a serious “challenge to our understanding” (Fellbaum, 2007) even when used in someone’s own mother tongue (L1). Interpreters are exposed to this challenge twice, as they have to render them properly and within a few seconds from their L1 to L2/s (i.e. to their foreign language/s).

Mistakes in the interpretation process can put both the validity of the whole interpretation and the interpreter’s credibility at stake. Our studies examine MWU interpretation errors to determine if they are made either at the comprehension level in the source language (SL) and/or if they develop during the language production in the target language (TL). Furthermore, we investigate if errors depend on language direction (English [EN] ><German [GE]). Wordnets are applied as a tool to determine comprehension and production errors. Based on these studies, we propose an ontological schema for interpreting studies with wordnets as a learning and validation tool.

1 Background and Motivation

In this article, we aim to introduce WordNet and its applications to the emerging field of interpreting studies. Our studies focus on multi-word units (MWUs), which represent both a challenge and an important source of errors in translation, interpreting and in natural language processing. We analyze the way professional and student interpreters deal with formulaic language units, particularly collocations and phrasemes. We also discuss the way in which wordnets can be fully implemented in an ontological learning method.

Two studies are introduced in this paper. The first one examines EU interpreters’ performances in simultaneously interpreting MWUs. After establishing the distribution of errors, we tried to locate, in the second study, their origin. We ex-

amined interpreters’ note-taking and introduced wordnets as reference tools, in order to determine whether the errors arise during the comprehension of the SL or during the language production into the TL. The results were then compared to see if errors may differ according to language direction (GE><EN). Based on these two studies, we eventually proposed an ontological approach, anchored to wordnets, for interpreting studies.

2 Theoretical and Methodological Considerations

A definitional frame was required in order to analyze the extracted forms descriptively. The final, unapodistic formulated etiquette, MWUs, refers to both collocations and phrasemes in a wider sense. The umbrella term, MWUs of language use, summarizes the attempt to focus on a large-scale investigation of these expressions, more than that of trying to justify their complex linguistic nature.

Collocations are defined, in this context, as holistic lexical, lexico-grammatical and semantic units composed by at least two elements. One of their compositional word retains its original meaning. They have a variable degree of idiomativity, and some of them accept variable composing elements through synonymy, an aspect which makes them recurrent but not fixed. In both studies, collocations were separated into grammatical classes, including VN collocations, AN collocations, AdvV collocations and PP collocations. These grammatical components should not be considered fixed in their sentence position, as they have to suit both languages (i.e. German and English) despite their different morphosyntactic alignment.

The term ‘phraseme’ is used as superordinate term and category name for idiomatic phrases, multi-word units, kinograms, doublets, triplets, metaphors, onymic phrases, phraseological terms, truisms, sayings, winged words and routine forms

such as speech openings and endings.

The studies' results were first classified as 'right' or 'wrong'; thereafter, they were more precisely defined through subcategories. These categories serve as assessment parameters for the evaluation of professional and student interpreters' performance, presented as follows.

'Right' interpretation acts were differentiated into the following categories:

a) Complete normative renderings (cnr). All peculiar elements of a collocation or a phraseme are preserved in both SL and TL. The entire grammatical form and discourse intention meant within the unit are also rightly conveyed:

- (1) Wir werden sie in guter Erinnerung behalten.
⇒ "We will always have a good memory of her".

b) Partial normative renderings (pmr). The MWU and its meaning are preserved but a grammatical mistake occurred.

- (2) for those who worked hard ⇒ "für diejenige, die hart arbeiten" (for those who work hard)

c) Paraphrases (par) of one collocation in case of nested MWUs or simplifications of the whole unit's meaning by using different words, of which nevertheless at least one is the base or the collocator of the original collocation. This rendering does not compromise the correct translation of the original unit's meaning and gives ground to the thought that the interpreter understood and recognized MWUs in the SL, but was not able to render them properly with equivalent forms in the TL.

- (3) Er hat eine Lücke hinterlassen, die nur schwer zu erschließen sein wird ⇒ "It has left an enormous gap to fill", instead of e.g. "He has left an enormous gap that will be difficult to fill."

- (4) The microphone will be cut off ⇒ "Der Mikrophon wird abgestellt". A more idiomatic form could be: "Der Mikrophon wird abgeschaltet".

d) Semantic chalks (sc). The MWU in the TL 'copies' the syntactic layout of the MWU in the SL. The meaning of the multi-word string can still be clearly understood, but the TL's expression is not as idiomatic as it should be.

- (5) Wir wollen ein sehr breites Zielpublikum ansprechen ⇒ "We want to address a broad public", instead of e.g. "We want to address a wide audience."

'Wrong' MWUs were categorised into two

main groups.

a) Omissions, i. e. SL units, which have been partially or completely left out in the TL. This affects the understanding of both the overall meaning and the meaning of the discourse where the MWU lies.

- (6) Woe betide us if we jeopardise freedom! ⇒ – [The MWU was not translated.]

b) Distortions (dist.). Wrong translation of the SL unit through a self-made unit (not idiomatic, not recurrent); adding of another, arguable information, not delivered in the original text, which distorts the whole discourse's meaning or adoption of a word allegedly considered to be the right synonym, which nonetheless changes the sentence (or discourse) information.

- (7) Ich verurteile das Verbrechen auf das Schärfste ⇒ "I utterly condemn the crime" ('Utterly' is a *quantitative* evaluation; 'auf das Schärfste', on the contrary, a *qualitative* one. A possible rendering could be: "I condemn the crime in the strongest terms.")

- (8) Higher costs ⇒ "Kosten in die Höhe treiben" (the MWU "in die Höhe treiben" approximately means 'to boost expenses', while the speaker meant 'höhere Kosten.'

In both studies, we tried to measure the performance of the interpreters on the base of these evaluation parameters.

3 Evaluation of EU Interpreters' Performance

The first preparatory study aims at investigating and evaluating the performance of EU interpreters, ten people for language combination, in rendering MWUs in their own mother tongue. They worked simultaneously in the booth and translated opening speeches, keynote addresses, reports to the EU Parliament and requests to speak, which we refer to as 'speeches' with general content and no jargon.

We studied a total of 113 GE>EN speeches and 147 EN>GE speeches. The total amount of GE>EN collocations and phrasemes was 1,243, while the sample for EN>GE amounted to 1,985 samples. All the original speeches were firstly transcribed, then linguistically analyzed¹. The linguistic research on transcripts of EU *interpreters'* performances was approved by the current EU Di-

¹Speeches of EU representatives are individually downloadable since September 1st, 2008.

GE>EN VN collocations						
✓: 81%			x: 19%			
cnr	pnr	par.	—	dist.		
85%	0%	15%	33%	67%		
GE>EN NA collocations						
✓: 80%			x: 20%			
cnr	pnr	par.	—	dist.		
50%	50%	0%	100%	0%		
GE>EN PP collocations						
✓: 88%			x: 12%			
cnr	pnr	par.	—	dist.		
73%	5%	22%	67%	33%		
GE>EN phrasemes						
✓: 77%			x: 23%			
cnr	pnr	par.	—	dist.		
68%	3%	29%	67%	33%		
EN>GE VN collocations						
✓: 86%			x: 14%			
cnr	pnr	par.	—	dist.		
58%	25%	17%	50%	50%		
EN>GE AdvV collocations						
✓: 83%			x: 17%			
cnr	pnr	par.	—	dist.		
60%	0%	40%	0%	100%		
EN>GE PP collocations						
✓: 83%			x: 17%			
cnr	pnr	par.	—	dist.		
87%	7%	6%	100%	0%		
EN>GE phrasemes						
✓: 90%			x: 10%			
cnr	pnr	par.	—	dist.		
60%	5%	35%	40%	60%		

Table 1: Evaluation of EU interpreters' performance - most significant results

rectorate General for Interpretation and Conferences, the current Head of the EU Plenary Records Unit and the current EU Data Protection Officer of the European Parliament, under the condition that the transcriptions would remain unpublished². The results were evaluated according to a set of criteria, as a golden standard is not available. In interpreting studies, it is on one hand very difficult (if not impossible) to collect sufficient data from a homogeneous interpreters' population with the same language combination. On the other hand, it is often unattainable to determine how a utterance should be segmented into smaller units (and whether there is a one-to-one mapping of these units in the SL and the TL). Moreover, it is quite impossible to identify how a single unit is best rendered (as demonstrated by Pöchhacker [2011] and Albl-Mikasa [2008]). In our study, the data were evaluated by an experienced interpreter after careful comparison of the transcriptions and after consulting known language resources in both TL and SL.

Table 1 and 2 below only report the most significant results. The correct MWUs are represented by a checkmark (✓), the wrong ones with an 'x'. All subcategories are abbreviated with the given acronyms, except for omissions, that are indicated with a dash (—). The data collected from the first study only refer to mistakes made while translating *into* the interpreters' mother tongue.

Phrasemes, VN and PP collocations were particularly difficult to render (posed special problems) for both English and German interpreters (as visible in Table 1). The most frequent strategy noticed was *anticipation*, i. e. the act of anticipating the speaker's said MWU: right phrasemes were

anticipated by English interpreters in 67 percent of the cases and by their German colleagues in 75 percent of the cases.

Nevertheless, the first study's results did not allow understanding at which level of comprehension and in which language direction mistakes mostly occurred (whether in the input or in the output phase). Were mistakes made because interpreters did not understand and/or know MWUs in the SL, or rather because they missed an unambiguous mental representation of their meanings, independently from their working language? We tried to find this out in the second study, by testing student interpreters' note-taking (consecutive interpreting) in both their L1 and L2.

4 Evaluation of Student Interpreters' Rendering of MWUs

4.1 Note-Taking: A Brief Introduction

In order to understand the proceeding of the second study and its outcome, we need to make a brief introduction on note-taking, the technique used to test student interpreters.

Note-taking or consecutive interpreting can be fully considered to be a parameter to investigate **mental representations of meaning/s**.

It is not the same as shorthand. With the latter, we generally refer to partial alphabetic systems; note-taking is, on the contrary, an interlingua (Fellbaum 2011) combining ideograms, logograms, abbreviations, letters of the alphabet, icons, symbols and some signs taken from i. a. stenography and mathematics. Note-taking resembles Blissymbolics³, because it represents concepts rather than words. Symbols can therefore stand for glosses, synsets, frames and entire

²E-mail exchange: November 29th, 2010.

³<http://www.alysion.org/handy/althandwriting.htm>

GE>EN VN collocations							EN>GE VN collocations						
✓: 92%				x: 8%			✓: 95%				x: 5%		
cnr	pnr	sc	par.	—	dist.	cnr	pnr	sc	par.	—	dist.		
GE>EN phrasemes													
✓: 73%				x: 27%			✓: 82%				x: 18%		
cnr	pnr	sc	par.	—	dist.	cnr	pnr	sc	par.	—	dist.		
52%	8%	32%	8%	0%	100%	83%	5%	8%	4%	77%	23%		
56%	13%	18%	13%	83%	17%	8%	0%	15%	77%	75%	25%		

Table 2: Evaluation of students' performance - most significant results

propositions⁴. Notes are generally ‘overlapped’ when they refer to the same concept⁵. Little changes are made on signs to disambiguate them from the context⁶, while the morpho-syntactic variations of a word are generally added on the upper or lower, right or left side of the root-note⁷. Signs could therefore excel in simplicity. From an ontological view point, they span on a note pad like *Top Ontologies* (TO) and Upper Domain Ontologies (UDOs), which stand for and represent both a root as well as its synset/s. Despite these few structural agreements however, one major unresolved problem in the note-taking system is arbitrariness. There are currently different opinions and theories on the most suitable signs and symbols to adopt, their preciseness and associated meaning, their amount, the note language used for abbreviations and the general content structure. This randomness, apparently justified by a solialistic approach to the system (with the purpose of immediate communication), leads to a frequent use of new symbols just “for the occasion”. This, as a result, remarkably affects the preciseness and general textual coherence of the rendering. It also partly justifies the difficulties in defining sound and objective evaluation criteria and quality parameters in consecutive interpreting.

Phrasemes and collocations are notably diffi-

⁴ ↗n means e.g. ‘a quantity that is added’, ‘to increase, augment, rise’, and, when combined with little more signs, e.g. ⇐ρ⇒¶, ‘to enhance, strengthen, reinforce the political dialogue’. The same applies e.g. to x/sth.↓, which means ‘to give an issue more weight, to lay more emphasis, to lend greater weight to sth.’).

⁵This applies to the case of ‘international automotive in-

dstry’, commonly represented as: . Also visualized overlapped MWUs should nonetheless be clearly readable.

⁶A ‘child’ is e.g. represented on a smaller scale; a ‘woman’ can wear a “skirt (△)”.

⁷The adjective ‘ugly’ is e.g. stressed by a minus to distinguish it from ‘horrible’. The adjective ‘visible’ can be drawn by using the same ideogram for ‘eye’, as it pertains to the word, however ‘visible’ is put on the upper right side of the noun or verb it refers to.

cult to notate in consecutive interpreting: according to Albl-Mikasa (2008: 256), the base is generally represented, while the collocate is omitted (which might lead to possible mistakes in the rendering).

4.2 The study’s outcome

The study was conducted during the preparatory session for an upcoming medical conference on a population of ten undergraduate German student interpreters in their last semester of study. They were asked to listen to 30 short texts of medical content in German and English, to which MWUs were purposely added (for a total of 70 MWUs). Interpreters were then asked to notate MWUs and render them in L1 or L2. They were encouraged to integrate WORDNET SEARCH – 3.1.⁸, WORDVIS⁹, the EN><GE online dictionary LANGUA.DE¹⁰ ^{‡‡} and other reference tools (i.a. EU’s multilingual term base IATE¹¹) to find right MWUs or check if they were correct. Through the study, we could indirectly test if wordnets extensively report MWUs in a clearly readable form and with little search.

The most significant results in the interpreters’ performance (reported in Table 2) were noticed in both language directions in **phrasemes and VN collocations**. Interpreters seemed to have less problems in rendering MWUs into their mother tongue than in translating them into English: more EN>GE phrasemes were in fact correctly translated than GE>EN phrasemes (82 percent against 73 percent of correct units). The same applies to VN collocations (95 percent right EN>GE VN collocations against 92 percent right GE>EN VN MWUs). Interestingly, all wrong GE>EN VN

⁸<http://wordnetweb.princeton.edu/perl/webwn>

⁹<http://wordvis.com/>

¹⁰<http://langua.de/>

^{‡‡}The second and third tools are respectively a WORDNET visual interface and a WORDNET aligned dictionary. For more information visit: <http://wordnet.princeton.edu/wordnet/related-projects/>.

¹¹<http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load>

MWUs were distorted. The majority of the interpreters recognized the respective bases of the MWUs, but had some problems in rendering the entire multi-word strings in both language directions (GE><EN). The difficulty did not seem to lie at the note-taking level; in fact, interpreters noted the known MWUs (even if differently) and could correctly define their meanings. This is an interesting aspect, as mistakes in the output phase of a consecutive session are generally attributed to an unclear or lacking note-taking system, which was clearly not the case in this instance.

5 WordNet and the Ontological Notation in Interpreting

Our studies confirmed that MWUs are indeed a major source of errors in interpreting and that errors, be in the SL comprehension or in the TL production, can be linked to the lack of an easily accessible mapping of the single conceptual meaning of a MWU. What fails is in other words a pivot between the idiosyncratic mapping of multi-word strings in SLs and TLs. Our studies also showed that wordnets can be powerful tools by providing the ontological link to the meaning of the units. However, when consulting wordnets, interpreters found that while some MWUs are clearly explicated, others are not recorded. One possible solution to help overcome the “resistance” to learn collocations and phrasemes and/or the difficulty to visualize them, could consist in entering MWUs in wordnets as single entries. This would enhance the ability to anticipate them while interpreting. In order to develop tools tailored to interpreters’ needs, other factors should also be taken into account: multilinguality should e.g. be contemplated in wordnet templates. We suspect that this would remarkably increase the search speed. A clear overview of the entries could be maintained by combining trimming and pruning as suggested by Huang (2010), and by applying saliency as presented by Vossen (2001). Moreover, it could be useful for interpreters to be given a visual representation of the meaning next to an entry. The adoption of a standardized note-taking system and of one frame of notation should be therefore implied. Other tools¹² are already using fragmentary word notation to improve learning and memory effectiveness; we could further exploit the potential of non-linear notation by developing an ontol-

ogy of language-independent consecutive notes. Wordnets, already serving as dictionaries and thesauri, could eventually further expand to editorial dictionaries, which would provide interpreters with general knowledge patterns for a concept.

References

- Michaela Albl-Mikasa. 2007. *Notationssprache und Notizentext. Ein kognitiv-linguistisches Modell für das Konsekutivdolmetschen*. Günter Narr Verlag, Tübingen.
- Aldo Gangemi *et al.*. 2003. “Sweetening WORDNET with DOLCE” in *AI Magazine*. Volume 24 No 3: 13-24.
- Daniel Gile 1995. *Regards sur la recherche en interprétation de conférence*. Presses universitaires de Lille, Lille.
- Chu-Ren Huang *et al.*. 2010. *Ontology and the Lexicon. A Natural Language Processing Perspective*. Cambridge University Press. New York.
- Helmuth Feilke 1996. *Sprache als soziale Gestalt. Ausdruck, Prägung und die Ordnung der sprachlichen Typik*. Suhrkamp Verlag. Frankfurt am Main.
- Christiane Fellbaum. 2009. *Sprachliches Wissen und Weltwissen: Wie Wörter und Begriffe zusammenhängen*. Georg-August-Universität Göttingen. <http://www.youtube.com/watch?v=aWvgTdS-koY>
- Christiane Fellbaum *et al.*. 2007. *Idioms and Collocations. Corpus-based Linguistic and Lexicographic Studies*. Continuum. London.
- Roberto Navigli. 2002. *Automatically Extending, Pruning and Trimming. General Purpose Ontologies*. Proceedings of the International Conference on Systems, Man and Cybernetics. ieeexplore.ieee.org/iel5/8325/26342/01168049.pdf?arnumber=1168049
- Franz Pöchhacker. 2011. “Assessing aptitude for interpreting. The SynCloze test.” *Interpreting* 13:1, pp. 106-120. John Benjamins Publishing Company.
- Princeton University “About WordNet.” WordNet. Princeton University. 2010. <http://wordnet.princeton.edu>
- Danica Seleskovitch; Marianne Lederer. 1996. “A systematic approach to teaching interpretation” in *Interpreting 1:1*, pp. 139-141. John Benjamins Publishing.
- Dirk Siepmann. 2007. “Wortschatz und Grammatik: Zusammenbringen, was zusammengehört”, *Beiträge zur Fremdsprachenvermittlung* 46, pp. 59-80. http://www.vep-landau.de/bzf/2007_46/bzf_Heft_46-2007.htm
- Piek Vossen. 2001. *Extending, trimming and fusing WordNet for technical documents*. The Association for Computational Linguistics. Proceedings on NAACL-2001 workshop on WordNet and other lexical resources applications, extensions and customizations. Pittsburgh, USA.

¹²E.g. relfinder.dbpedia.org/ www.smartwisdom.com/

Low-cost ontology development

Marek Grác and Adam Rambousek

NLP Center

Faculty of Informatics, Masaryk University

Botanická 68a, 602 00 Brno

Czech Republic

{xgrac, xrambous}@fi.muni.cz

Abstract

In this paper, we present the project building new lexical resource – shallow ontology derived from the corpora. The ontology should be used primarily for machine translation, syntactic parsing and word sense disambiguation. Currently, the ontology for Czech language is developed, but the methodology and tools are suitable for other languages with similar structure.

Ontology is based on BushBank corpus, which improves handling of ambiguity in natural language. BushBank data and tools are application-driven, thus reducing the time and costs needed to annotate the corpora and develop new lexical resources.

1 Introduction

Language resources for natural language processing are very important for development as well as improvement of existing natural language processing (NLP) tools. Situation for different European languages varies a lot. In the worst case there are almost no resources and we have to face the problem of creating them cheaply and quickly while maintaining high quality. We can attempt to build an ultimate corpus that will be useful for every application but we do not believe that such approach is successful often enough. We have decided to model our corpus using application-driven development. This approach should prevent major design flaws which might not be automatically recoverable later and could limit the usefulness of resulting work for ours needs.

This approach was used to build a multi-layered annotated corpus which is one of the resources used for creating our ontology. Application-driven development means that at begining corpus does not contain any data directly usable for creating

ontology as we avoid creating data with no immediate application (even if it might be useful in future). In fact proposed ontology is independent of original corpus and we plan to use also other (larger) corpora for enriching ontology.

This paper focuses on a new type of annotated corpus named BushBank and an example study of building shallow ontology for Czech language on top of it. Semantic networks are among the most popular formalisms for knowledge representation. Like other networks, they consists of nodes and links. For English we have a number of possibilities from domain oriented to general ones like Princeton WordNet (Miller, 1990). It is very rich and complex network but unfortunately only few applications use its potential.

Creation of similar resources for other close languages such as Czech is very difficult and also time-consuming. Our goal is to create a simpler ontology which will be easy to create and use primarily in our existing applications. This application-driven approach should help us to avoid creating a perfect complex ontology by providing us with a simple one instead, which can be used in various projects right now. Simpler ontology should also help us to create similar resources for other languages and take advantage of it in machine translation. Such project can reuse many existing components that were created for different purposes and projects.

2 WordNet ontologies

A lexicon with information about how words are used and what they mean is a necessary component for any application working with natural language. Ontologies are one of the resources that can provide enough information for those. Ontology is a formal representation of a set of concepts within a domain and the relationship between those concepts. Ontologies can be based on different assumptions, for specific domains and

different purposes. Thus it is very difficult to compare them using objective metrics.

There are several ontologies built for the English language. For smaller European languages, one of the most important general ontologies is Princeton Wordnet (Miller, 1990). It contains many relations (e.g. hypo/hyperonym, is part of) connecting synsets (synonym set) which are equated with ‘senses’. Specifically, according to WordNet’s on-line glossary, a *sense* is a ‘a meaning of a word in WordNet. Each sense of a word is in a different synset’. Princeton WordNet is available under free license also for commercial applications.

EurowordNet (Vossen, 1998) and Balkanet (Christodoulakis, 2004) were projects to localize (and improve) parts of the original version to Central and South East European languages. Thanks to ILI (inter lingual index), it is possible to connect ontologies and use the result as a multilingual dictionary. Unfortunately some of the problems of original WordNet still remain (Hanks and Pustejovsky, 2005), e.g. the assumption that membership in two or more synsets is equivalent to having more different senses. Some of the WordNet senses are indistinguishable from one another by any criterion. Attempt to build a WordNet-like ontology for new language was described in various papers (Pala and Smrž, 2004; Erjavec and Fišer, 2006). Creation of proper synsets and assigning the relations is a time-consuming process that needs expert in this field. One of the most serious problems of the EWN data is their very strict license.

3 VerbaLex

VerbaLex is the lexicon of verb valencies for Czech language, developed at the Faculty of Informatics, Masaryk University. VerbaLex (Hlaváčková and Horák, 2005) combines valency frames and formalism, used in previous projects (Balkanet and Vallex 1.0 (Hajic et al., 2003)), with other relevant information, such as verb aspect, verb synonymy and semantic verb classes based on VerbNet project (Schuler, 2005). VerbaLex contains 10 478 verbs, 21 123 verb senses and 19 360 valency frames. Information in VerbaLex is stored in the form of *complex valency frames* (CVF).

Complex valency frame is designed as a sequence of elements which form a list of necessary

grammatical features (e.g. preposition and grammatical case).

opustit:4/leave office:1 (give up or retire from a position)
frame:

AG <person:1>^{obl}_{whol} VERB ACT<job:1>^{obl}_{what4}
example:

Jarek opustil zaměstnání / Jarek left his job

Example sentence can show us that if “Jarek” has to be the agent (semantic role) then it has to be in nominative (numbered 1) case. Also it has to be a hyponym of person:1 in the WordNet ontology. Thanks to ILI we can have nodes named in English and use words from Czech EuroWordNet.

This notation exported to an XML format allows us to easily process both syntactic and semantic layer of the sentence.

4 Annotation process

Annotation of linguistic data is considered to be a task for experts. This is especially right for those corpora that attempt to cover more layers or structures of a language. Process of annotation is usually described in detail in an annotation manual. As an example, we can take annotation manual for the semantic layer of PDT2.0 which spans tens of pages (Hajič et al., 2005). In last years, we have witnessed several attempts to use crowdsourcing for small parts of linguistic annotation (Munro et al., 2010).

In order to use crowdsourcing we have to find a crowd that exceeds a critical mass. Thanks to services like Amazon Mechanical Turk, this is usually not a problem for widely used languages, such as English. Situation for languages like Czech (10 million speakers) is more complicated as no services of this type are available.

We have decided to involve students. Our annotators are mostly in their first year at the university and they have very limited amount of deeper linguistic knowledge. Our previous experience with student annotators gives us some hope that they can be trained to carry out simple linguistic tasks better than an average crowd-member, though.

We assume that an annotation standard is usually an attempt to approximate several mutually exclusive and contradictory constraints (Jakubíček et al., 2010):

1. **completeness:** the annotation should provide

complete linguistic insight into the particular area;

2. **consistency**: the annotation should be consistent, i.e. same or similar language phenomena should be handled in same or similar ways;
3. **usability**: the annotation should enable straightforward usage in the intended applications;
4. **simplicity**: the annotation should be as simple as possible to make high inter-annotator agreement achievable.

In our experience, most language resources try to find a trade-off among the constraints by prioritizing them in the order given above. They prefer completeness over consistency, and both of them over simplicity.

Following the so-called KISS¹ principle, we are strongly convinced that the reverse order of those constraints represents a much better priority list to be met when building a language resource. Thus, our priorities are:

- **simplicity**: so that annotators do not err too often;
- **usability**: so that the usage of the resource will be straightforward;
- **consistency**: following from simplicity;
- **completeness**: just in case everything is simple, usable and consistent.

Main objection against this new order of priorities can be that consistency is crucial to most NLP application. This applies to using the data both for testing/development and for machine learning. From our perspective, natural language and its semantics is too ambiguous and flexible to be easily and consistently annotated. We have to face situations where even expert human annotator encounters a possibility of having more than one correct annotation. Inconsistencies between annotators are traditionally resolved by an expert who decides which annotation is correct. Qualified opinion of an expert can improve consistency of annotations but we do not prefer to use other than crowdsourcing methods. As we would like to know that

these examples are clear and others are ambiguous for annotators. This can help us to better test applications as we can't expect machines to handle semantic ambiguity better than people and thus testing should be performed mostly on clear cases.

We had attempt to constrain the annotators as much as possible with a simple annotation scheme. Annotators can not add new noun phrases (nodes in ontology) and they have to work only with noun phrases found in source material. As can be seen on screenshot annotators can answer only yes/no. Limiting creativity and working with preprocessed data helps us to increase inter-annotator agreement and (therefore) also consistency.

5 BushBank corpus

BushBank is a concept that extends TreeBanks, which are sets of annotated syntactic trees, by reducing the requirements for unambiguity and making them closer to real language. Like other modern corpora, bushbank usually covers several layers of linguistic annotation. For this reason, we have decided to use NXT NITE (Carletta et al., 2005), which was developed for multimodal corpora. We do not plan to have a multimodal corpus, but using existing libraries for complex search queries and the XML format persuaded us. On top of this toolkit, we have built our own library which maps elements in the corpus to objects, so that programmers do not need to care about the internal NXT NITE structures or about XML elements.

One of our main objections against existing annotated corpora is the fact that they treat language as an unambiguous structure and possible ambiguities are solved by the annotation manual or by expert decision. This leads to a situation when corpora users are not able to determine whether they are handling cases that were easy to determine or cases where even human annotators were not really sure. For various NLP applications, it is crucial to know whether the application can handle correctly at least the clear cases and only later focus on areas which are hard even for humans.

Ambiguity in a BushBank is one of its main advantages. In fact, only the first layer has to be disambiguated. This layer contains marks for sentences and a token for every word in the corpus. We are aware that even on this layer, it is possible to have ambiguities but both simplicity and usability will be corrupted if we introduce ambiguity at

¹Keep It Simple and Stupid

this level. Currently for the Czech BushBank (as first case-study of bushbank concept) we have the following layers:

1. **tokens**: contains tokens and marks for begin/end of sentence.
2. **morphology**: defines lemma and morphologic tag for tokens.
3. **syntactic structures**: defines short noun phrases, verb phrases, coordination and clauses. This structures uses the token layer.
4. **relations between syntactic structures**: for every short noun phrase we define its dependency parent.

We believe that corpus users should be able to select proper resolution model for their needs and thus they should have access to the existing annotation also in the form of raw data. All our results are easily reproducible and can be reproduced by anyone interested in doing so.

6 Building the Sholva ontology

The Sholva ontology, currently being developed at FI MU, attempts to create a new lexical resource for Czech language which will be free to use for any purpose. Proposed methods and implementation of the tools are also suitable for other languages with similar structure (e.g. Slovak) and we believe that those languages will follow soon. Our ontology should be used primarily for machine translation, syntactic parsing and word sense disambiguation. Application driven extensions should be possible.

We do not intend to create an ontology with dozens of relations and complicated. For our usage, only hypo/hyperonymical relations can be used directly, and basic ontology will not contain any other relations. In EWN, senses of words are numbered, but splitting word senses is also a very subjective task. More importantly, this has no direct relevance to our primary goals. For word sense disambiguation, we need to be able to distinguish various senses of a given word. We believe that for this purpose, the knowledge of path from the root node to any given word created by hyponymical relations defines *word sense*. It is possible that a word will have several hyperonymical relations, but we do not know if they refer to the same or a different thing.

The process of creating an ontology in this style is very similar to corpora annotation. Annotation tool provides users with very simple interface. Annotation tool is web-based and optimized for mobile devices, see screenshots 1 and 2. Each annotator is given a set of around 500 tokens per annotation round. This set can be annotated on iPad device in 5–6 minutes as annotator can only answer by clicking/touching *yes* or *no* buttons.

Each token consists of semantic class (e.g. person:1) and noun phrase for which annotator have to select its validity. Annotator do not have possibility to look at whole original sentence or found examples in corpus. This may look like a step backwards as language resource build on top of corpus are used for quite long time. But it is not. Our noun phrases are from corpus but we want ontology that do not contains figurative language. For figurative language context of noun phrase is very important so we have decided to preventively ignore context at all.

Candidate tokens for annotation are prepared by combination two layers from bushbank: syntactic structures (noun and verb phrases) and relations of syntactic structures and verb valency lexicon. Using bushbank helps us to find only those noun phrases which are mapped to verb as valencies. These noun phrases are then mapped to valency frames of given verb. As each of the slot of VerbaLex valency lexicon points to node in WordNet we are able to suggest potential semantic class for noun phrase.

In later stages of building ontology, we plan to use existing ontology to improve precision of our mapping by using only those valency frames where at least one of the slot will be filled by noun phrase which is already known and can be in expected semantic class.

7 Conclusion

Application driven development that lead us to creation of new ontology means that ontology was created to solve our specific problems. Our main idea is to use this information to better disambiguate mapping of verb valency frames and to test current recall and precision. We expect that successfull mapping should helps us to improve machine translation dictionary as we will be able to translate valency frames instead of verbs. Additionally this mapping can be used to improve anaphora resolutions.

In the future, the project will continue with the annotation of additional resources and we plan to develop methods to use also large corpora that are not annotated in as detailed way as our bushbank. We plan to release our corpus to the research community. Along with that, our linguistic tools and resources will be improved by fixing problems discovered in the process of annotation. We will gladly help to create a similar resource for other languages. We believe that this can be a way for even smaller languages to obtain valuable linguistic resources, using this very low-cost approach.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and LINDAT-Clarin project LM2010013, by the Czech Science Foundation under the projects P401/10/0792 and 102/09/1842, and by the Ministry of the Interior of CR within the project VF20102014003.

References

- J. Carletta, S. Evert, U. Heid, and J. Kilgour. 2005. The NITE XML toolkit: data model and query language. *Language resources and evaluation*, 39(4):313–334.
- D. Christodoulakis. 2004. *Balkanet Final Report*. University of Patras, DBLAB. No. IST-2000-29388.
- T. Erjavec and D. Fišer. 2006. Building slovene wordnet. In *Proceedings of the 5th International Conference on Language Resources and Evaluation LREC*, volume 6, page 24.
- J. Hajic, J. Panevová, Z. Urešová, A. Bémová, V. Kolárová, and P. Pajas. 2003. Pdt-vallex: Creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9, pages 57–68.
- J. Hajíč, J. Panevová, E. Buráňová, Z. Urešová, A. Bémová, J. Kárník, J. štěpánek, and P. Pajas. 2005. Anotace na analytické rovině. *Návod pro anotátory*. Praha: ÚFaL MFF UK.
- J. Hajic. 1998. Building a syntactically annotated corpus: The prague dependency treebank. *Issues of valency and meaning*, pages 106–132.
- P. Hanks and J. Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10(2):63–82.
- Dana Hlaváčková and Aleš Horák. 2005. Verbalex – new comprehensive lexicon of verb valencies for czech. In *Proceedings of the Slovko Conference*, Bratislava, Slovakia.
- M. Jakubíček, V. Kovář, and M. Grác. 2010. Through low-cost annotation to reliable parsing evaluation. In *PACLIC 24 Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 555–562.
- G. Miller. 1990. Five Papers on WordNet. *International Journal of Lexicography*, 3(4). Special Issue.
- R. Munro, S. Bethard, V. Kuperman, V.T. Lai, R. Melnick, C. Potts, T. Schnoebelen, and H. Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 122–130. Association for Computational Linguistics.
- Karel Pala and Pavel Smrž. 2004. Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 7(1-2):79–88.
- K.K. Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.
- P. Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer.

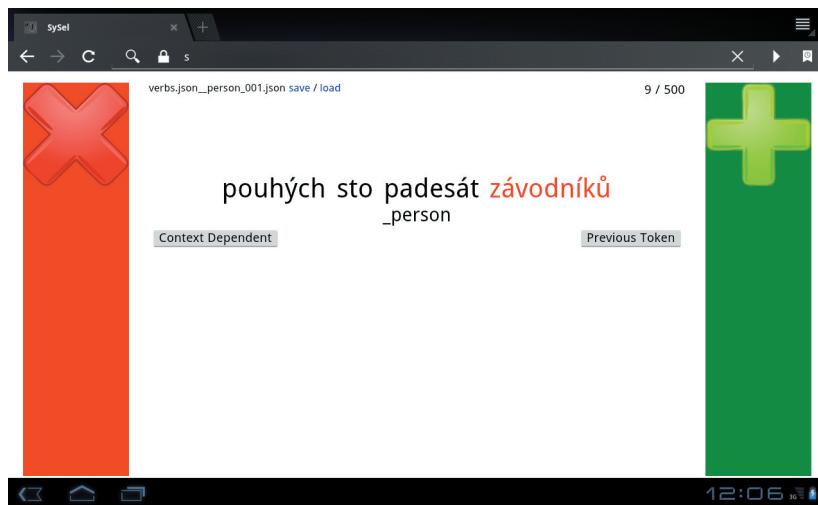


Figure 1: Annotation tool on Android tablet

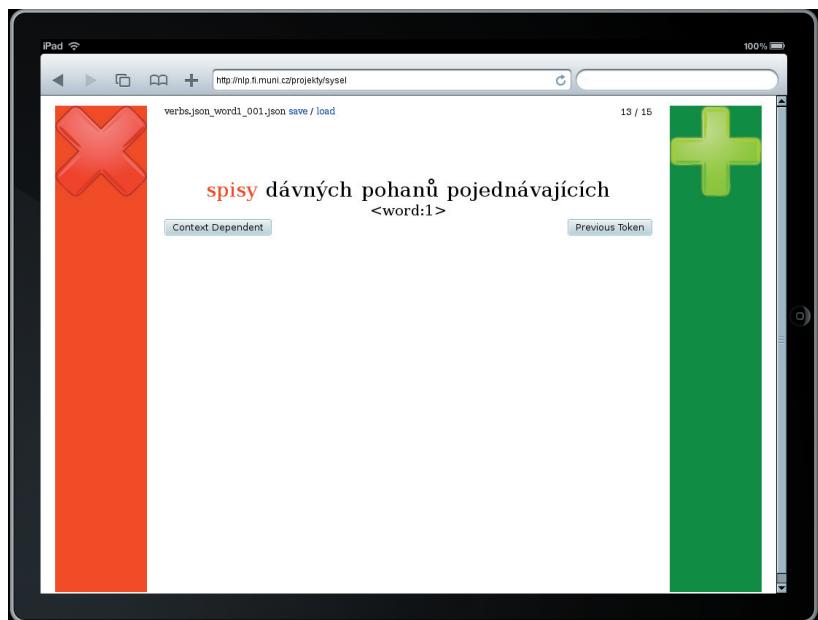


Figure 2: Annotation tool on iPad

Migrating Cornetto Lexicon to New XML Database Engine

Aleš Horák and Adam Rambořík

NLP Center

Faculty of Informatics, Masaryk University

Botanická 68a, 602 00 Brno

Czech Republic

{hales, xrambous}@fi.muni.cz

<http://deb.fi.muni.cz>

Abstract

The original Cornetto project started to develop a new complex-structured lexicon for the Dutch language. The lexicon building process works with information from two current electronic dictionaries – the Referentie Bestand Nederlands (RBN), which contains FrameNet-like structures, and the Dutch wordnet (DWN) with the usual wordnet structures. The resulting Cornetto lexicon is stored in a system called Cornetto database, which is built over the Dictionary Editor and Browser platform.

In this paper, we describe a transition of the Cornetto database system to a new database backend based on large set of tests that were run on four selected (out of twenty) available XML database systems. We present the technical details of the Cornetto editing process and the results before and after the database transition.

1 Introduction

The Cornetto database system (Horák et al., 2009) is based on the DEB (Dictionary Editor and Browser) development platform (Horák et al., 2006a). The general purpose of DEB is to offer common client-server functionality for different types of lexicographic resources, including dictionaries, wordnet semantic networks, classical ontologies or lexical databases.

The Cornetto lexico-semantic database¹ combines Wordnet with FrameNet-like information (Fillmore et al., 2004) for the Dutch language. During the lexicon building process the Dutch Wordnet (Vossen, 1998) and the Referentie Bestand Nederlands (Maks et al., 1999) are the

most consulted external language resources. The Dutch Wordnet (DWN) is similar to the Princeton Wordnet for English, and the Referentie Bestand Nederlands (RBN) includes frame-like information as in FrameNet plus additional information on the combinatoric behaviour of words in a particular meaning. The combination of the two lexical resources results in a rich linguistic database that improves natural language processing (NLP) technologies, such as word sense-disambiguation, and language-generation systems. In addition to merging the Wordnet and FrameNet-like information, the database is also mapped to a formal ontology to provide a more solid semantic backbone.

Both DWN and RBN are semantically based lexical resources. RBN uses a traditional structure of form-meaning pairs, so-called Lexical Units (Cruse, 1986). Lexical Units contain all the necessary linguistic knowledge that is needed to properly use the word in a language. Lexical Units in Cornetto are organized into Synsets (synonymical sets). For Cornetto, the Synsets follow the concept of near synonymy from EuroWordNet (Vossen, 1998).

The DEB platform is currently employed in more than 15 national and international projects, e.g. the KYOTO EU project (Vossen, 2008) or a five-year project of the New Encyclopaedia of the Czech Language. Two projects with nearly thousand active users all over the world are DEBDict and DEBVisDic (Horák et al., 2006b). DEBDict as a general dictionary browser offers access to many dictionaries and lexical resources in several languages, and DEBVisDic, wordnet editor and browser has already been used to build more than fifteen wordnets in different languages from all over the world. The freely available DEB server is currently installed in ten institutions from three continents, where it is used mostly as an XML-based data storage, presentation and manipulation system.

¹see Figure 1

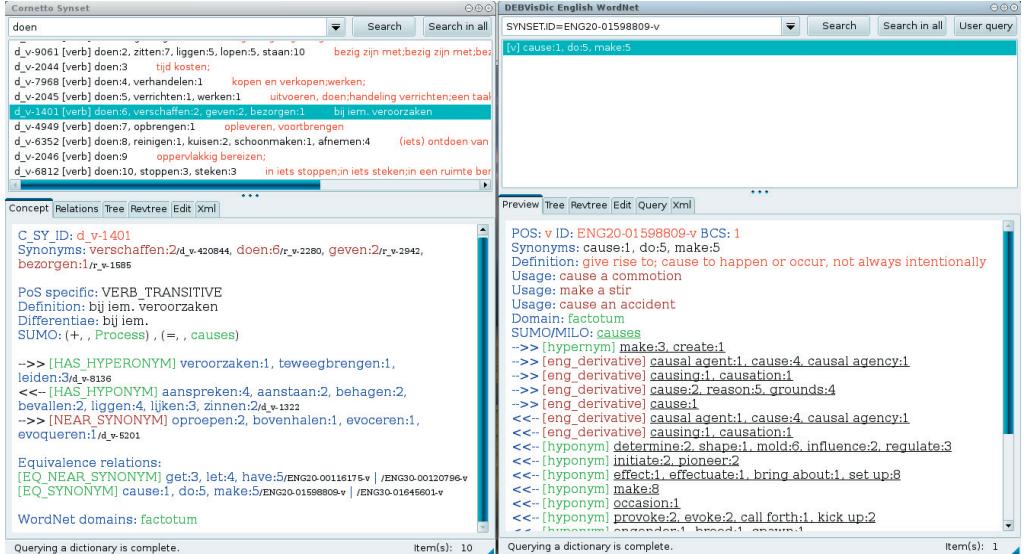


Figure 1: An example of the Cornetto editing interface.

In the following text, we first describe the needs of a DEB storage backend and the selection process between possible XML databases used as a storage backend in the DEB platform, and then we present the details of changing the backend within the Cornetto database system.

2 The DEB Database Backend

With the current deployment of the DEB platform, several complex tasks have appeared with growing needs for the employed database storage. To resolve such problems and to offer reserves in the speed of the DEB database backend (i.e. the XML storage engine used for saving the data processed by the system), available native XML database systems were analyzed and compared, with the resulting recommendation of the best performance for knowledge and ontology systems.

The DEB (Dictionary Editor and Browser²) is an open-source software platform for the development of applications for viewing, creating, editing and authoring of electronic and printed dictionaries. The platform is based on the client-server architecture. Most of the functionality is provided by the server side, and the client side offers (computationally simple) graphical interfaces to users. The client applications communicate

²<http://deb.fi.muni.cz/>, see e.g. (Horák et al., 2006a)

with the server using the standard web HTTP protocol.

The server part is built from small, reusable parts, called servlets, which allow a modular composition of all services. Each servlet provides different functionality such as database access, dictionary search, morphological analysis or a connection to corpora.

The overall design of the DEB platform focusses on modularity. The data stored in a DEB server can use any kind of structural database and combine the results in answers to user queries without the need to use specific query languages for each data source. The main data storage is currently provided by the Oracle Berkeley DB XML (Chaudhri et al., 2003). However, it is possible to switch to another database backend easily, without any changes to the client parts of the applications.

Database systems working with XML data (both native XML databases and XML enabled relational databases) are already widespread and used in many areas. Their performance was benchmarked by many projects using several benchmarks. However, conclusions of previous publications (Böhme and Rahm, 2008; Nambiar et al., 2002; Lu et al., 2005) do not provide one definitive answer as for the choice of the best XML database. XML-enabled and native XML

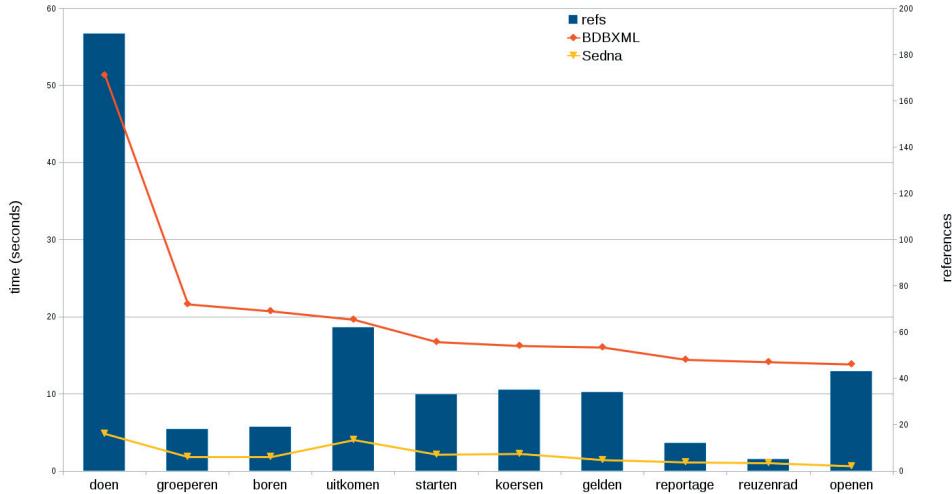


Figure 2: Graph comparison of the most time-consuming Cornetto synset queries in Oracle Berkeley DB XML and Sedna

databases. Generally, the results suggest that different XML benchmarks can show different weak and strong points of each database systems. When comparing the two classes of XML databases, i.e. relational databases with XML support and native XML databases, we can see that XML enabled relational databases process data manipulation queries more efficiently, and native XML databases are faster in navigational queries which rely on the document structure.

We have thus performed extensive testing with the selection of 4 from more than 20 native XML or XML-enabled databases. The selection was driven by the requirements of effective XML processing, an open source licence, active development and support of XML-related standards. The tested databases were:

- *eXist*. The eXist database (Meier and others, 2003) is developed in Java and licensed under LGPL, active since 2000 and currently developed by the group of independent developers. The database supports XQuery, XSLT and XUpdate standards for data manipulation, and DTD, XML Schema, RelaxNG and Schematron for validation.

Users are able to specify structural indexes (element and attribute structure in docu-

ments), range indexes (*contains*, *starts-with* and similar functions), and full-text indexes (Apache Lucene (Foundation, 2006) is used for full-text indexing).

- *MonetDB*. The MonetDB database (Boncz et al., 2006) is developed by CWI Amsterdam and several Linux distributions and MS Windows are officially supported. The database is licensed under a customised Mozilla Public License.

The main goal of MonetDB is to design a database for processing very large (in GBs) XML documents. The default database settings are optimized for document reading, offering indexing for quick query execution, although the indexes have to be rebuilt after every document update. Another option is an optimization for document updating, with simpler index structure and slower performance for search queries.

The database supports XQuery and partly XQuery Update (W3C, 2009). It is also possible to use MonetDB internal query language. Indexing is automatized, without the possibility to alter settings in any way. The PF/Tijah (Hiemstra et al., 2006) text search system is utilized for full-text searching.

Table 1: Complex synset searches in Oracle Berkeley DB XML and Sedna (in seconds)

The slowest queries				in Sedna		
	# of refs	DBXML	Sedna		# of refs	Sedna
doen	189	51.4	4.9	proper	31	8.3
groeperen	18	21.7	1.9	omlaaglopen	12	7.9
boren	19	20.8	1.9	gaan	216	7.0
uitkomen	62	19.7	4.1	houden	143	7.0
starten	33	16.8	2.2	zin	145	6.8
koersen	35	16.3	2.3	stuk	168	6.8
gelden	34	16.1	1.5	oppassen	44	6.7
reportage	12	14.5	1.2	hol	58	6.6
reuzenrad	5	14.2	1.1	hand	43	6.5
openen	43	13.9	0.7	slag	117	6.3

The most frequent searches				in Sedna				
	# of searches	# of refs	DB XML	Sedna	# of searches	# of refs	Sedna	
god	594	24	2.7	1.8	artikel	128	53	1.2
artikel	550	53	4.7	1.2	aardig	127	62	1.4
jacht	337	57	0.9	0.5	intreden	120	17	0.9
schijf	265	36	1.3	2.2	gewoonte	113	26	1.1
officier	260	75	0.9	2.0	komen	77	146	0.6
arm	252	61	1.6	3.1	krijgen	73	31	1.1
college	214	65	1.1	2.1	vallen	64	115	1.4
academie	206	30	0.8	1.5	slaan	62	73	0.5
rijkdom	197	38	0.9	2.0	inbrengen	60	21	0.9
president	194	9	0.5	0.5	heer	58	138	4

- **Sedna.** The Sedna database system (Fomichev et al., 2006) is developed by the Russian Academy of Sciences, and released under Apache Licence. Official packages for Windows, Linux, MacOS, FreeBSD and Solaris are available.

The database supports XQuery and custom variant of XQuery Update for data manipulation, and XML Schema for validation. Indexes have to be set manually and a special function must be used in the query to access the index. Full-text indexing is provided by external commercial tool dtSearch. Sedna offers several extensions, such as the capability of an SQL connection from XQuery, or the trigger support.

- **Oracle Berkeley DB XML.** Oracle Berkeley DB XML (Chaudhri et al., 2003) was created as an extension of Berkeley DB. The database

is now developed by Oracle and released for Windows and Linux. Users can choose between open source and commercial licences.

The underlying structure is still based on Berkeley DB and each document container is stored in a single file. The database supports XQuery and part of XQuery Update. The document validation according to a supplied XML Schema is checked only during document storage, later changes can render the document invalid. Users have to specify indexes manually, full-text indexing is also supported, although it is not possible to use regular expressions in queries.

Because of the special focus on dictionary writing systems, we ran different test suites designated to both “raw speed” of the database and to specific requirements of knowledge and ontology systems. According to the results of the tests (see (Bukatović et al., 2010) for the details of the tests re-

Table 2: Lexical unit search in Oracle Berkeley DB XML and Sedna (in seconds)

The slowest queries				
in Oracle Berkeley DB XML		in Sedna		
	DBXML	Sedna	Sedna	
uitkomen	25.3	0.5	sterk	9.1
vervallen	22.9	0.4	doteren	6.6
steken	21.2	1.0	prioriteit	4.5
opstaan	20.9	0.8	gelijk	4.3
trekken	20.7	0.9	aanvaarding	4.2
opzetten	20.7	0.5	zwaar	4.2
sterven	20.6	0.8	onbeschaafd	3.9
treden	20.5	0.5	vurig	3.7
plaatsen	20.5	0.4	open	3.3
springen	20.5	0.3	onmogelijk	3.3

The most frequent searches						
in Oracle Berkeley DB XML		in Sedna				
	# of searches	DBXML	Sedna	# of searches	Sedna	
god	560	1.8	0.6	schilderen	1173	0.1
artikel	533	3.2	0.2	draaien	520	0.8
eindy	183	1.0	1.0	slaan	453	0.8
gewoonte	152	5.4	0.3	gebruiken	413	0.1
vis	143	0.4	0.6	keren	349	0.5
gang	127	1.4	1.1	branden	343	0.5
richtlijn	123	0.2	0.7	verliezen	317	0.2
beeld	114	1.4	0.3	blazen	306	0.2
huis	102	1.3	0.7	hechten	294	0.3
goed	101	0.9	1.4	steunen	279	0.2

sults), none of the available native XML databases can supersede the others for all kinds of operations needed for knowledge and ontology storage and manipulation. Berkeley DB XML cannot efficiently solve the queries involving multiple nodes and full-text queries. The eXist database contains the Lucene module for text search and supports many XML standards, so it can be recommended for deployment where these features are more important than the database performance. On the other hand the MonetDB database can be, according to its specific architecture, conveniently used for when working with very large amounts of XML data. For middle-size data collections, the Sedna database can provide the same performance as MonetDB, while offering richer set of features. The potential drawbacks of Sedna are the need to use special queries for the defined data indexes and the use of commercial tool for optimized full-text queries. However, the full-text queries without this optimization are already comparably fast.

During the testing of both database engines within the DEB platform, we found out that the MonetDB programming interface for the Ruby language used in the DEB platform is not stable enough and not developed actively at the moment. Because of that, MonetDB is not ready yet to be included in the platform. Fortunately, Ruby interface for Sedna is stable and maintained and better suited for DEB platform. That is why Sedna was chosen for the DEB database backend transition for one of the very active projects with tens of concurrent editors, the Cornetto project.

3 Cornetto Backend Transition and Evaluation

The Cornetto data are split to four main databases (lexical units, synsets, ontology terms, and Cornetto identifiers), plus two databases with more detailed linguistic information and the English wordnet database. Usually, user queries combine data from several databases and different informa-

tion are merged to form complex entries.

Database sizes and number of links between them grew over time, currently the main databases contain:

- lexical units – 117 967 entries;
- synsets – 70 507 entries;
- identifiers – 106 305 entries;
- ontology – 3 080 entries.

As the database size and complexity increased, search queries were getting slower even with indexes set up to speed up the most common queries in the current database backend Oracle Berkeley DB XML. Because the user experience was an issue, Cornetto was chosen as the first project to migrate to new database backend.

The database module of DEB platform was exchanged from Oracle Berkeley DB XML to Sedna and no explicit indexes were set up. Even without the indexes, the improvement in search queries was considerable. We have analyzed the database logs for both implementations. The speed was measured in the same conditions (machine, workload, number of users and their interaction with the software). Logs for Berkeley DB XML cover the usage from April 2010 to March 2011. Logs for Sedna cover the usage from March 2011 to September 2011. All presented times are averages of all the searches for the respective entry words and represent just the time needed to run the query in database and to prepare the result list, not the time of the client-server communication over the Internet.

Tables 1 present query times for complex database queries on synsets, with references to lexical units, English Wordnet and ontology. The tables show top 10 slowest queries and 10 most frequent queries in both DB XML and Sedna. For the DB XML words the corresponding times in Sedna are also displayed for comparison. It is clear that the search times are affected by the number of references each synset contains. For example, there are 10 synsets for the word *doen*, with 146 references to other synsets, 28 references to lexical units and 15 references to the English Wordnet. Similar

Tables 2 show statistics for lexical units. The tables present top 10 slowest queries and 10 most frequent queries in Oracle Berkeley DB XML and Sedna.

Table 3: Average time to execute queries in Cornetto database (in seconds)

	Berkeley DB XML	Sedna
Synset	6.1	2.4
Lexical units	2.7	0.9

Finally, Table 3 summarizes average search times (in seconds) for Berkeley DB XML and Sedna databases.

Considering the results of XMark and the custom knowledge and ontology benchmark, the MonetDB and the Sedna databases represent a good choice for the knowledge and ontology systems. MonetDB offers very good performance for very large documents, on the other hand, Sedna provides much more advanced features. Sedna supports index usage only with its own special functions, so the queries need to be changed accordingly.

According to the experiences and evaluation of the Sedna database deployment for Cornetto, performance improvement is significant and enhances user experience. Database performance will be enhanced even more by utilizing specific indexes to speed up the query execution. Based on this pilot transition, the Sedna database will be included in the DEB platform and will gradually replace Oracle Berkeley DB XML as the main database backend.

Based on the preliminary tests, the performance is also greatly affected by the XML parser library included in the DEB platform. Currently, REXML (Russell, 2008) parser is used for parsing each entry during search. However, other parser libraries can improve the speed significantly. Evaluation of available XML parsers will be carried out to find the best option for DEB platform.

4 Conclusions

In the paper, we have presented a successful adaptation of the Cornetto database system to a new XML database backend. The Cornetto system is based on the Dictionary Editor and Browser (DEB) development platform that is designed to provide modular framework for dictionary writing systems. The structure of the Cornetto system and the DEB platform thus allows to change the underlying data storage without the need to make substantial changes to the system as a whole.

We have described the details of the database

selection process and the evaluation of the database transition. The presented results, as well as positive user experience, clearly justify that the new database is very well suited to the kind of operations needed for the development of the Cornetto Lexicon as a new complex lexico-semantic language resource.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and LINDAT-Clarin project LM2010013, by the Czech Science Foundation under the projects P401/10/0792 and 102/09/1842, and by the Ministry of the Interior of CR within the project VF20102014003.

References

- T. Böhme and E. Rahm. 2008. Multi-user evaluation of XML data management systems with XMach-1. *Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web*, pages 148–159.
- P. Boncz, T. Grust, M. van Keulen, S. Manegold, J. Rittinger, and J. Teubner. 2006. MonetDB/XQuery: a fast XQuery processor powered by a relational engine. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, page 490. ACM.
- M. Bukatovič, A. Horák, and A. Rambousek. 2010. Which XML storage for knowledge and ontology systems? In *Knowledge-Based and Intelligent Information and Engineering Systems*, pages 432–441. Springer.
- Akmal B. Chaudhri, Awais Rashid, and Roberto Zicari, editors. 2003. *XML Data Management: Native XML and XML-Enabled Database Systems*. Addison Wesley Professional.
- D.A. Cruse. 1986. *Lexical semantics*. Cambridge, England: University Press.
- C.J. Fillmore, C.F. Baker, and H. Sato. 2004. Framenet as a 'net'. In *Proceedings of Language Resources and Evaluation Conference (LREC 04)*, volume vol. 4, 1091–1094, Lisbon. ELRA.
- A. Fomichev, M. Grinev, and S. Kuznetsov. 2006. Sedna: A Native XML DBMS. *Lecture Notes in Computer Science*, 3831:272.
- Apache Software Foundation. 2006. Apache Lucene. <http://lucene.apache.org>.
- D. Hiemstra, H. Rode, R. van Os, and J. Flokstra. 2006. PF/Tijah: text search in an XML database system. In *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR)*, pages 12–17.
- A. Horák, K. Pala, A. Rambousek, and M. Povolný. 2006a. DEBVisDic—First Version of New Client-Server Wordnet Browsing and Editing Tool. In *Proceedings of the Third International WordNet Conference (GWC 2006)*, pages 325–328, Jeju Island, South Korea.
- Aleš Horák, Karel Pala, Adam Rambousek, and Pavel Rychlý. 2006b. New clients for dictionary writing on the DEB platform. In *DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writings Systems*, pages 17–23, Italy. Lexical Computing Ltd., U.K.
- A. Horák, I. Maks, A. Rambousek, R. Segers, H. van der Vliet, and P. Vossen. 2009. Cornetto Tools and Methodology for Interlinking Lexical Units, Synsets and Ontology. In *Current Issues in Unity and Diversity of Languages*, pages 2695–2713, Seoul, Republic of Korea. The Linguistic Society of Korea.
- H. Lu, J.X. Yu, G. Wang, S. Zheng, H. Jiang, G. Yu, and A. Zhou. 2005. What makes the differences: benchmarking XML database implementations. *ACM Transactions on Internet Technology (TOIT)*, 5(1):154–194.
- I. Maks, W. Martin, and H. de Meerseman, 1999. *RBN Manual*.
- W. Meier et al. 2003. eXist: An open source native XML database. *Lecture Notes in Computer Science*, pages 169–183.
- U. Nambiar, Z. Lacroix, S. Bressan, M. Lee, and Y. Li. 2002. Efficient XML data management: an analysis. *E-Commerce and Web Technologies*, pages 261–266.
- Sean Russell. 2008. Rexml. <http://www.germane-software.com/software/rexml/>.
- P. Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer.
- Piek Vossen. 2008. KYOTO Project (ICT-211423), Knowledge Yielding Ontologies for Transition-based Organization. <http://www.kyoto-project.eu/>.
- W3C. 2009. XQuery Update Facility 1.0. (<http://www.w3.org/TR/xquery-update-10>).

A Proposed Nepali Synset Entry and Extraction Tool

Arindam Roy
Assam University
Silchar
arindam_roy74@rediffmail.com

Sunita Sarkar
Assam University
Silchar
sunitasarkar@rediffmail.com

Bipul Shyam Purkayastha
Assam University
Silchar
bipul_sh@hotmail.com

Abstract

Lexically rich resources form the foundation of all NLP tasks. Maintaining the quality of resources is thus a high priority issue. In this paper we propose a tool for the purpose of inserting lexemes/words constituting a synset and a few other important characteristics of a synset into the database and then looking up the synset using a synset extraction tool. The tool has been developed for Nepali Wordnet

Correspondingly a number of tools were developed to provide better functionality and transparency to this impeccable language resource which not only forms the heart of all Wordnets in India, but also of all NLP work in India.

2 ROAD MAP

The Road Map of the paper is as follows:- Section 3 provides an overview of some special features of Nepali language, Section 4 provides a description of the Nepali WordNet , Section 5 describes the tool for synset insertion and look up and Section 6 concludes the paper.

3 Features of Nepali Language

The hereditary structure of Nepali language is:- Indo European>Indo Iranian>Indo Aryan>North Western>Khasa Prakrit>Pahadi Language>Eastern Pahadi(Nepali). Nepali, like Hindi and its ancestor Sanskrit , unlike English, is a *Subject Object Verb* (SOV) language, i.e., in Nepali, the subject, object, and verb of a sentence usually appear in that order. For example:-

Sentence: उसले मेरो केरा खायो।

Transliteration: usle mero keraa khaayo.

Gloss: he my banana ate.

Parts: Subject Object Verb

Translation: He ate my banana.

Nepali is written in Devnagari script. Nepali is a Head-right language i.e. in every phrase the head is on the right. The typical order of a VP is NP-VP. The typical order of a NP is ADJ-NP. The typical order of ADJP is ADV-ADJP.

1 Introduction

Princeton English WordNet (C. Fellbaum, 1998) is an ontological, machine readable lexical database for English language developed at Princeton University. It graduated to become one of the most used and prized among language resources. When a language resource of quality as high as the English WordNet came into being, several tools were developed to utilize, enhance and maintain the resource as best as possible.

Since the birth of the English WordNet, WordNets for many other languages have spawned. In case of Indian languages, Hindi WordNet (Dipak Narayan et al., 2002) was the first of its kind. Hindi Wordnet was developed at the IIT Bombay.

¹Nepali is a language in the Indo-Aryan branch of the Indo-European family. It is the official language of Nepal and in India it is the country's one of the 23 official languages. The *Ethnologue* website counts more than 17 million speakers worldwide. In India there are about 5,00,000 Nepali speakers in Sikkim while in Darjeeling and Jalpaiguri districts of West Bengal there are about 1,400,000 speakers of Nepali.

The typical order of PP is NP-PP i.e. the language is postpositional. It is written phonetically, that is, the sounds correspond almost exactly to the written letters. Nepali has many loanwords from Arabic and Persian languages, as well as some Hindi and English borrowings. There are some deviating features of Nepali among the Indo-Aryan languages. These are:-

a. Unknown Past Aspect:- Let us take the sentence ‘*kukhuro marechha*’ (Nepali). A loose translation in English is ‘*chicken is dead*’. But actually it should be ‘*chicken was found to be dead*’(at the time the speaker came to know of this fact). ‘*chicken is dead*’ has an equivalent Nepali translation ‘*kukhuro maryo*’. But as the *death* has occurred in some unknown past, the Nepali speakers tend to say ‘*kukhuro marechha*’.

b. Gender:- Human genders are treated as masculine or feminine. Apart from humans, all other nouns are treated as masculine. For e.g. ‘*Ram aayo*’(Ram came), ‘*Sita aai*’(Sita came).

‘*Goru aayo*’(Ox came), ‘*Gaii aayo*’(cow came)

c. Number:- There are two numbers in Nepali- *eka bachan* and *bahu bachan*. ‘*Haru*’ is the plural marker. But it is not imperative to use the plural marker for all nouns. For e.g. *keto aayo*(boy came,singular), *Ketaa aaye*(boys came,plural). Here ‘*o*’ ending nouns change into ‘*aa*’ ending for plural sense and verb ‘*aayo*’ becomes ‘*aaye*’ for plural sense. As a result, sentences can be made plural without using plural marker.

4 NEPALI WORDNET

Nepali Wordnet is a system for bringing together different lexical and semantic relations between the Nepali words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. The design of the Nepali WordNet is based on the principle of “expansion” from the Hindi Wordnet and English Wordnet. This principle was first proposed within the EuroWordNet project (Vossen, 2002). Thereafter it has been used by a number of WordNet development teams for the creation of new WordNets. Examples include the WordNets for Spanish, French (Vossen, 2002), Hungarian language (Alexin et al., 2006), Hindi, Marathi (Sinha et al., 2006) etc. In the Expansion Approach, synsets of a preexisting WordNet are understood by the lexicographer and the

corresponding target language synsets expressing the same sense are created.

4.1 FEATURES OF NEPALI WORDNET

In Nepali Wordnet, the words are grouped together according to their similarity of meanings. Two words that can be interchanged in a context are synonymous in that context. For each word there is a synonym set, or synset, in the Nepali WordNet, representing one lexical concept. This is done to remove ambiguity in cases where a single word has multiple meanings. Synsets are the basic building blocks of WordNet. The Nepali WordNet deals with the content words, or open class category of words. Thus, the Nepali WordNet contains the following category of words- Noun, Verb, Adjective and Adverb.

Each entry in the Nepali Synset consist of the following elements:-

ID: The synset identifier.

CAT: The syntactic category of the sense.

CONCEPT: It explains the concept represented by the synset. For example, “*यस्तो कुरा वा काम जसले कसैको मान वा प्रतिष्ठा कम गराउँछ*” (*yastokuraa waa kaam jasle kasaiko maan waa pratishTha kam garaaūcha*) explains the concept of insult as some saying or deed which diminishes somebody’s reputation.

EXAMPLE: It gives the usage of the words of the synsets in the sentence. In general, the words in a synset are replaceable in the sentence. For example: “*हामीले कसैलाई पिन अपमान गनुहुँदैन*” (*haameele kasailaaee pani apmaan garnuhūdain*) gives the usage for the words in the synset of ‘अपमान’, ‘apmaan’ representing insult as something that should not be done to anybody.

Synsets in Nepali have been classified as:- Universal,Pan Indian,Language Specific, Synthesised and Rare.

a. Universal Synset:- Linkable across all languages of the world with natural and indigenous lexemes/words(e.g. the synset of “sun”)

ID :: 2186

CAT :: noun

CONCEPT :: हाम्रो सौर जगतको

त्यो सबैभन्दा ठुलो र ज्वलनशील पिण्ड

जसद्वारा सबै ग्रहहरू तातो र उज्यालो हुन्छन्

EXAMPLE :: सूर्य सौर उर्जाको
एउटा धैरे ठुलो स्रोत हो /पूर्वमा सूर्य उदाएको
देखेर अँध्यारो पुच्छर लुकाएर भागयो

SYNSET-NEPALI :: सूर्य, धाम, रवि

b. **Pan Indian**:- Linkable across all languages of India with natural and preferably indigenous lexemes/words. (e.g. "papad"; a kind of crispy food)

ID :: 6171
CAT :: noun
CONCEPT :: चना, मुङ्ग, चामल
आदिको पीठोलाई मसला लगाएर पातलो बेत्तेर
सुकाएको खाद्यवस्तु जसलाई सेकेर वा फुराएर
खाइन्छ

EXAMPLE :: आमा पापड सेकै
हुनुहन्छ

SYNSET-NEPALI :: पापड

c. **In-Family**: Linkable across all Indian languages belonging to a family (Indo-Aryan, Dravidian and Sino-Tibetan) with natural and preferably indigenous lexemes/words. (e.g. "bhatijaa" meaning brother's son , "mama" meaning mother's brother, "kaka" meaning father's brother .These are all naturally occurring expression in Indo Aryan family of languages)

ID :: 9783
CAT :: noun
CONCEPT :: आमाको भाई
EXAMPLE :: किरणका मामा

उच्च न्यायालयमा वकील छन्

SYNSET-NEPALI :: मामा

ID :: 7379
CAT :: noun
CONCEPT :: भाइको छोरो
EXAMPLE :: आज श्यामको

भतिजो आँनेछ

SYNSET-NEPALI :: भतिजो, भतिज

c. **Language specific**:-No issue of linkage here. Only synsets expressing concepts that are common in a specific language have to be carefully included in the synset repository. Other languages can link to them only by constructing artificial phrases or through

transliteration. It is imperative to allocate synset id spaces to these synsets like in Unicode. For example, {पेवा [pewaa, a portion of the property of family owned by a female meber]} is a sense in Nepali which does not have any corresponding sense in Hindi.

- i. Transliteration
 - ii. Use of multiword expression (short phrases)
 - iii. Coining of new words
- The steps should be used in the above order of priority.

d. **Synthesised**:- The words used by all languages in same sense/meaning . For e.g. rickshaw.

e. **Rare**:- These are synsets expressing rare concepts. (e.g. highly technical terms).These will necessarily be transliterated and a range bearing very large id numbers will be allocated for these.

5 Synset Entry and Extraction Tool

To facilitate a simple synset insertion point for the linguists, the Synset entry tool was developed. This synset data entry interface tool has been developed in Java and is linked with an MS-Access database at the back-end. The data entry interface tool has features for entering the ID of a particular synset, the category of the synset(i.e. the POS of the synset) , the gloss which expound how the principal word of the synset can be used in a sentence and finally the words that represent the synset.

Once the synsets are entered in the database, they can be searched or extracted using a tool which has been developed using ASP and Java Script. The ASP script runs in the IIS server. The user types in the head-word of the synset which he/she wants to search in a text box using a virtual keyboard . The virtual keyboard has been developed using JSP. It has keys which represent all the alphabets of the Nepali language and 'yuktaakshars' can be typed using this virtual keyboard. The virtual keyboard enables the user to directly type the lexeme/word which he/she intends to search. He/She does not need to, at first, look up any Nepali keyboard map which in itself is a cumbersome and time taking job and then type the lexeme/word for look-up. The virtual

keyboard uses Unicode for the various Nepali alphabets and symbols.

The lexemes/words typed by an user is the index to the database. An ASP script running in a Microsoft IIS server searches the database for the synset and if the synset is present, then it returns the ID of the synset, the category of the synset, the gloss of the synset and the words constituting the synset

The words in a synset may have different senses if their categories are different. For example

ID	:: 20630
CAT	:: NOUN
CONCEPT	:: त्यो विधान जो
मानिएका मूलभूत सत्यहरूको आधारमा सिद्ध गरिन्छ	
EXAMPLE	:: "उनको शोधकार्यसित सम्बन्धित १२० वटा प्रमेय छन्"
SYNSET-NEPALI:: प्रमेय	
ID	:: 20629
CAT	:: ADJECTIVE
CONCEPT	:: जुन प्रमाणको विषय हुन सक्छ
EXAMPLE	:: "प्रमेय वस्तुहरूलाई जम्मा गरेर राख्न आवश्यक छ"
SYNSET-NEPALI:: प्रमेय	

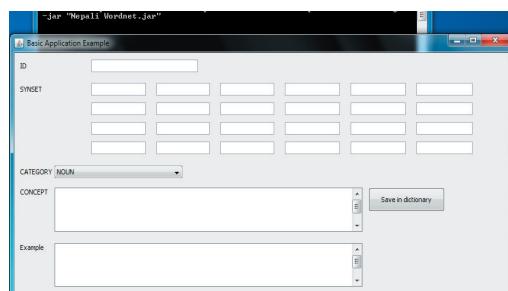


Figure 1:A tool for inserting synsets into the database



Fig 2: Synset Extraction Tool

Figure 3 : Result of Synset extraction tool

6 Conclusion

The synset entry tool has been developed keeping in mind that lexicographers and linguists would need such type of tool while entering the synsets into the database. As the Nepali Wordnet is in the process of development so an ontology for the synsets, various relations like hyponymy-hypernymy,meronymy-holonymy, entailment, troponymy etc are yet to be developed in the wordnet.Cross part of speech linkage also needs to be developed. The future version of the tool would include the above mentioned features.

The synset extraction tool has also been developed keeping in mind that in future, the Nepali Wordnet would be put online just like the Hindi Wordnet. The purpose is that people can use it and put forward their suggestions, if any.

References

- [1] C.Felbaum:Wordnet : *An Electronic Lexical Database* , MIT Press,1998
- [2] Dipak Narayan, Debasri Chakrabarty, Prabhakar Pande and P. Bhattacharyya: *An Experience in Building the Indo WordNet- a WordNet for Hindi*, International Conference on Global WordNet (GWC 02), Mysore, India, January, 2002.
- [3] Kuhoo Gupta, Manish Shrivastava, Smriti Singh and Pushpak Bhattacharyya, *Morphological*

- Richness Offsets Resource Poverty- an Experience in Building a POS Tagger for Hindi*, COLING/ACL-2006, Sydney, Australia, July, 2006
- [4] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. *Five Papers on WordNet*. MIT press.1993
- [5] George A. Miller. 1995. English WordNet -A Lexical Database for English". *Communications of the association for Computing Machinery*, 38(11):39-41. 1995
- [6] Manish Sinha, Mahesh Reddy, Pushpak Bhattacharyya. *An Approach towards Constructionand Application of Multilingual Indo-WordNet*, 3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea, January, 2006
- [7] Zoltán Alexin, János Csirik, András Kocsor, and Márton Miháltz. *Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction*, 3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea, January, 2006.
- [8] Ganesh Ramakrishnan and Pushpak Bhattacharyya. *Word Sense Disambiguation Using Semantic Sets Based on the WordNet*. International Conference for Language Resource Evaluation: Special Workshop on Using Semantics for Information Retrieval and Filtering, Canary Islands, June, 2001.
- [9] Nepali language. 2009.
http://en.wikipedia.org/wiki/Nepali_language
- [10] Online Hindi WordNet. 2009.
<http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php>
- [11] Colin P. Masica. *The Indo-Aryan Languages*. Cambridge University Press, Cambridge, UK.1991
- [12] Alok chakraborty, Bipul Shyam Purkayastha, Arindam Roy, *Experiences in building the Nepali Wordnet* Proceedings of the 5th Global WordNet Conference, Mumbai, Narosa Publishing House, India.
- [13] Vossen P. (ed.). *EuroWordNet: A MultilingualDatabase with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.1998
- [14] S. Jha, D. Narayan, P. Pande, P. Bhattacharyya, *A WordNet for Hindi*, International Workshop on Lexical resources in Natural Language Processing , Hyderebad , India , January , 2001.
- [15] Bhattacharyya P., Fellbaum C. and Vossen P. (eds) *Principles, Construction and Application of Multilingual Wordnets*, Proceedings of the 5th Global WordNet Conference, Mumbai, Narosa Publishing House, India.
- [16] Arindam Chatterjee et al. *Tools for IndoWordnet and WSD* , Proceedings of the 5th Global WordNet Conference, Mumbai,Narosa Publishing House , India

Automatic Extension of WOLF

Benoît Sagot

INRIA / University Paris 7

Paris, France

benoit.sagot@inria.fr

Darja Fišer

University of Ljubljana

Ljubljana, Slovenia

darja.fiser@ff.uni-lj.si

Abstract

In this paper we present the extension of WOLF, a freely available, automatically created wordnet for French, the biggest drawback of which has until now been the lack of general concepts that are typically expressed with highly polysemous vocabulary that is on the one hand the most valuable for applications in human language technologies but also the most difficult to add to wordnet accurately with automatic methods on the other. Using a set of features, we train a Maximum Entropy classifier on the existing core wordnet to be able to assign appropriate synset ids to new words, extracted from multiple, multilingual sources of lexical knowledge, such as Wiktionaries, Wikipedias and corpora. Automatic and manual evaluation shows high coverage as well as high quality of the resulting lexico-semantic repository of. Another important advantage of the approach is that it is fully automatic and language-independent and could therefore be applied to any other language still lacking a wordnet.

1 Introduction

Whatever the framework and representation of lexical knowledge, such as ontologies, framenets or wordnets, semantic lexicons can only contribute to applications in human language technologies, such as word-sense disambiguation, information extraction or machine translation, if their coverage is comprehensive as well as accurate. The language resources development community seems to have reached a consensus that, despite giving the most reliable results, manual construction of lexical resources is to time-consuming and expensive to be practical for most purposes. Several semi- or fully automatic approaches have been proposed instead, exploiting various types of existing resources to facilitate the development of a new semantic lexicon, especially

wordnets. However, most proposed approaches to induce a wordnet automatically, still suffer from the necessary trade-off between limited coverage and the desired level of accuracy, both of which are required if the resource is to be useful in a practical application.

This is why we present here an approach for wordnet extension by extracting additional lexico-semantic information from already available bilingual language resources and then training a maximum entropy classifier on the existing core wordnet in order to assign the new vocabulary to the appropriate synsets. Our approach, applied on the French wordnet WOLF, is comprehensive in that it can handle monosemous and polysemous words from all parts of speech which belong to the general vocabulary as well as specialized domains and can also deal with multi-word expressions.

The rest of the paper is structured as follows: in Section 2 we give an overview of related work. In Section 3 we introduce the current edition of WOLF. In Section 4, we describe the process of extracting lexico-semantic information from bilingual lexical resources. In Section 5 we explain the wordnet enrichment experiment using a maximum entropy classifier that helped us determine whether a translation we extracted from the existing resources is an appropriate candidate for a given synset. Section 6 is dedicated to the analysis and evaluation of the extended resources, and Section 7 contains concluding remarks and ideas for future work.

2 Related work

Most automatic approaches to create a wordnet for a new language take the Princeton wordnet as a backbone and extend it with the vocabulary inventory of the target language. One of the most straightforward and most widely used resources to obtain lexical knowledge for the language in question are machine-readable bilingual diction-

aries. Entries from the dictionary are linked to PWN synsets under the assumption that their counterparts in the target language correspond to the same synset (Knight and Luk 1994). A well-known problem with this approach is that bilingual dictionaries are generally not concept-based but follow traditional lexicographic principles, which is why the biggest obstacle is the disambiguation of dictionary entries.

When such dictionaries are not available or when they do not contain sufficient information to disambiguate the entries, bilingual lexicons can be extracted from parallel corpora (Fung 1995). The underlying assumption here is that senses of ambiguous words in one language are often translated into distinct words in another language (Dyvik 2002). Furthermore, if two or more words are translated into the same word in another language, then they often share some element of meaning (Ide et al. 2002). This results in sense distinctions of a polysemous source word or yields synonym sets.

The third set of approaches that have become popular in the past few years extract the meaning, translations and relationships between words in one or several languages from Wikipedia. New wordnets have been induced by using structural information to assign Wikipedia categories to WordNet (Ponzetto and Navigli 2009) or by extracting keywords from Wikipedia articles (Reiter et al. 2008). Vector-space models to map Wikipedia pages to Wordnet have been developed (Ruiz-Casado et al. (2005)). The most advanced approaches use Wikipedia and related projects, such as Wiktionary, to bootstrap wordnets for multiple languages (Melo and Weikum 2009, Navigli and Ponzetto 2010).

An unsupervised machine-learning approach has been used by Montazery and Faili (2011) to construct a wordnet for Persian. Their approach is similar to ours in the sense that they too combine translation candidates obtained from bilingual dictionaries and corpus-based contextual information for these candidates to establish links to Princeton WordNet synsets.

3 Wordnet Libre du Français (WOLF)

Previous work on the development of WOLF (Fišer and Sagot 2008) has focused on benefiting from available resources of three different types: general and domain-specific bilingual dictionaries, parallel corpora and Wiki resources (Wikipedia and Wiktionaries).

The core WOLF was created by disambiguating (*literal, synset*) pairs obtained from a word-aligned multilingual parallel corpus with the help of already existing wordnets for several languages other than French. For each multilingual lexicon entry, translation equivalents in all these languages were assigned a set of possible synset ids from their wordnets. Assuming that translation equivalents in the word-aligned parallel corpus are lexicalizations of the same concept, they shared one or several intersecting synset ids which were then also assigned to the French equivalent in the lexicon. The approach was limited to almost predominantly basic synsets which were common among all the wordnets used, and to single-word literals because cross-lingual mapping of multi-word expressions was not possible with or word-alignment procedure. In order to compensate these shortcomings, additional (*literal, synset*) pairs for monosemous English words were also harvested from various freely-available bilingual resources (dictionaries, thesauri and Wikipedia). Sense assignment for these was near perfect because they did not require any disambiguation.

The wordnet for French created in this way contained about 32,300 non-empty synsets, 87% of which were nominal. With the approach we adopted, we were able to populate just over 50% of Base Concept Sets and 25% of the rest of the concepts from Princeton WordNet. The first version of WOLF was already bigger than the French WordNet (22,121 synsets) that had been developed within the EuroWordNet Project (Vossen 1999) and was comparable to the more recent wordnet construction contribution called JAWS (34,367 synsets) which was developed by Mouton and de Chalendar (2010) from a bilingual dictionary, which however contains only nouns.

Manual evaluation of the results showed that the wordnet generated in this way is relatively reliable but does not use full potential of the available resources. This is why we have devised an additional large-scale extension cycle, aiming at taking full advantage of the existing lexical resources in order to improve the coverage of WOLF without compromising its accuracy while the first version of WOLF will serve as the baseline. The procedure is described in the rest of this paper. We begin by presenting the various resources used in the experiment and the way we extracted (*literal, synset*) candidates from them. Then we introduce the maximum entropy classifier and the features we use for filtering these

pairs and extending our initial wordnet. We also report manual and automatic evaluation of the results and look into possible steps to refine the developed resource in the future.

4 Bilingual lexicon extraction

In this experiment we used two types of sources of lexical knowledge: the structured freely-available general and domain specific bilingual dictionaries, and the semi-structured articles from the on-line Wikipedia. The main goal of the extraction process was to extract as many French translation variants for each English word as possible in order to capture as many senses of that word as possible. With this we obtained *wordnet candidates* in the form of (*literal, synset*) pairs, i.e. a French translation of an English word with an assigned synset id from Princeton WordNet.

General vocabulary was extracted from the English and French **Wiktionary** in which translations are explicitly encoded for all parts-of-speech. The number of pairs extracted from each resource is given in Table 1. For domain-specific vocabulary we used **Wikispecies**, a taxonomy of living species that includes both Latin standard names and vernacular terms.

Less structured than dictionaries but with a much more predefined structure than free text is the online multilingual collaborative encyclopaedia **Wikipedia**. We used English and French articles by following inter-language links that relate two articles on the same topic. We enhanced the extraction process with a simple analysis of article bodies with which we resolved ambiguities arising from the capitalization of article titles (e.g. *Grass-novelist*, *Grass-plant*). In a similar way we also identified synonyms for the article titles (e.g. “*Cannabis*, also known as *marijuana*”), their definitions (e.g. “*Hockey* is a family of sports in which two teams play against each other by trying to manoeuvre a ball or a puck into the opponent’s goal using a hockey stick.”) and usage examples.

Resources used	En-Fr equivalents
English Wiktionary	39,286
French Wiktionary	59,659
Wikispecies	48,046
Wikipedia	286,818
Total (duplicates removed)	417.419

Table 1: Results of bilingual lexicon extraction from heterogeneous resources

The result of our extraction process is a large bilingual lexicon of all English-French translation pairs with the name of the resource they originate from. The figures for both extracted bilingual lexicons are summarized in Table 1.

As can be seen in Table 1, we were able to extract a substantial amount of bilingual entries from the various resources. However, the extracted entries suffer from an important drawback: they do not contain any explicit information that can help us map these entries to PWN, neither do they contain contextual information from corpus occurrences that would help us determine their sense based on their usage. For example, an English-French translation pair (*dog, chien*), which we extracted from the Wiktionary, does not contain any information that would make it possible for us to determine which of the 8 synsets in WPN containing the literal *dog* would be appropriate to be translated with *chien* in WOLF. In Wiktionary articles, translations of a given word are sometimes organized into senses and described with short glosses. These have been compared to PWN glosses in order to map Wiktionary senses to PWN synsets (see Bernhard and Gurevych 2009). The first sentence of a Wikipedia article can be used in a similar way (see Ruiz-Casado et al. 2005). However, this is not the case for all Wiktionary entries or for other resources. Therefore, at this point, we assign to each translation pair *all* possible synset ids and disambiguate it in the next step.

5 Large-scale wordnet extension

Restricting the use of a bilingual lexicon to monosemous English literals is a safe but very limited approach that does not exploit the available resources to their full potential, which is a waste of resources and should be improved. However, using lexicon-based candidates generated from polysemous English literals is only possible if we can establish the likelihood with which a word should be added to a particular synset, i.e. can compute the semantic distance between a given French literal and PWN synset id.

In this paper we propose a technique to achieve exactly that. It is based on the core version of WOLF and a probabilistic classifier that uses various features associated with each (*literal, synset*) candidate.

5.1 Training set and feature selection

First we extract all (*literal, synset*) pairs from the bilingual lexicon that are already in the baseline wordnet and consider them as valid ones (score 1). All the other candidates, on the other hand, are considered invalid (score 0). This creates a noisy but reasonable training set for a probabilistic model. It is noisy for two reasons: first, our baseline wordnet does contain some mistakes because synsets were generated automatically and have not been completely manually validated; second, and more important reason is the fact that the baseline wordnet is not complete, which is why our new candidates may be valid even though they are not present in the baseline wordnets. It is precisely these candidates we are looking for in our wordnet extension procedure. In order to use the baseline wordnet as a training set for our classifier that will assign scores to all the candidates in the lexicon, we need to extract features from (*literal, synset*) pairs in WOLF.

5.1.1 Semantic proximity

The central feature we use models the **semantic proximity** between a literal and a synset. The feature can be illustrated on the example (*dog, chien*) we already used above. There are 8 PWN synsets that contain the literal *dog*, which is why this bilingual entry yields 8 different (*literal, synset*) candidates. We now need to determine which of these 8 candidates are valid. In other words, we need to establish which of the 8 corresponding synsets the French literal *chien* should be added to in WOLF. We therefore compute the semantic similarity of the literal *chien* w.r.t. each of these 8 synsets. For doing this, we first represent each WOLF synset by a bag of words obtained by extracting all literals from this synset and all the synsets up to 2 nodes apart in WOLF. For example, the synset *{andiron, firedog, dog, dog-iron}* in PWN, which is empty in the baseline WOLF, is represented by the bag of words *{appareil, mécanisme, barre, rayon, support, balustre,...}* (~device, mechanism, bar, shelfe, baluster,...). Next, we use a distributional semantic model for evaluating the semantic similarity of *chien* w.r.t. this bag of words. We use the freely-available SemanticVectors package (Widdows and Ferraro 2008). The distributional semantic model was built from the 65,000 lemmatised webpages from the French web corpus frWaC corpus (Ferraresi *et al.* 2010). This gives us a semantic similarity score between *chien* and the synset *{andiron, firedog, dog, dog-iron}*,

which is only 0.035, while the similarity between *chien* and one of its valid synsets, *{dog, domestic dog, Canis familiaris}* is as high as 0.331.

5.1.2 Additional features

In addition to semantic proximity, we use a number of other supporting features which are described below. Let us consider a candidate (T, S) that has been generated because our bilingual resources provided entries of the form $(E_1, T) \dots (E_n, T)$, where all PWN literals E_i 's are among S 's literals. **The number of such PWN literals** is one of the features. **Each possible source** (e.g. English Wiktionary) corresponds to one feature, which receives the value 1 if and only if at least one of the (E_i, T) entries was extracted from this source. We also extract **the lowest polysemy index** among all E_i 's: if one of the E_i 's is monosemous, this feature receives the value 1; if the least polysemous E_i is in two PWN synsets, this features receives the value 2. The idea is that if the candidate is generated from at least one monosemous PWN literal, it is very likely to be correct, whereas if it was generated from only highly polysemous PWN literals, it is much more questionable. Finally, **the number of tokens** in T is used as a feature (often, literals with many tokens are not translations of PWN literals but rather glosses).

5.2 Classifier training

Based on these features, we train a classifier using the Maximum-Entropy package *megam* (Hal Daumé III, 2004). An analysis of the models shows that the semantic similarity is by far the strongest feature. As expected, the lowest polysemy index among English literals also contributes positively, as does the number of different English literals yielding the generation of the candidate, and the number of sources involved. On the other hand, also as expected, the number of tokens in the target language literal has a negative impact on the certainty score.

The result of our classifier on a given (*literal, synset*) candidate is a score between 0 (bad candidate) and 1 (good candidate). We empirically set the threshold at 0.1 (see Section 6.1) for adding the candidate to the wordnet. The results are presented and evaluated in the next section.

6 Results and evaluation

6.1 Analysis of the results

Our wordnet extension procedure yielded 55,159 French wordnet candidates (out of 177,980).

Among the 55,159 French candidates, 15,313 (28%) correspond to (*literal, synset*) pairs already present in the previous version of WOLF, which means that 39,823 (72%) new pairs were added. As a consequence, 13,899 synsets that were empty in the previous version of WOLF now have at least one French literal.

A comparison of WOLF before and after the extension paralleled with the figures from Princeton WordNet 3.0 is given in Table 2. The extended version of WOLF has 43% more non-empty synsets than before the extension. The increase in the number of (*literal, synset*) pairs in the new WOLF is even higher; the number rose from 46,411 to 76,436 (+65%).

	PWN 3.0	WOLF old	WOLF new
N	82,114	28,559	36,933
V	13,767	1,554	4,105
Adj	18,156	1,562	4,282
Adv	3,621	871	1,125
Total	117,658	32,550	46,449
BCS1-3	4,671	4,339	6,171
Non-BCS	112,987	28,211	40,278

Table 2: Results of the wordnet extension procedure

As in PWN, by far the most frequent domain is Factotum, and the order for the following three most frequent domains is the same in both wordnets as well (Zoology, Botany, Biology). Most synsets belonging to these domains were generated from Wikispecies and Wikipedia while Wiktionary was the most frequent source for the Factotum domain. Of all the wordnet domains, only 3 are missing in WOLF (Paleontology, Rugby, and Volleyball) but these domains have less than 10 synsets in total even in PWN.

Average synset length in the extended WOLF is 1.79 literals per synset, which is slightly more than in PWN 3.0 (1.76). It is the lowest for nominal synsets (1.72) and the highest for adverbial ones (2.06). In PWN adverbial synsets are by far the shortest (1.54) while verbal ones are the longest (1.82). The longest synset in the extended WOLF is an adverbial one which contains as many as 27 literals, while in PWN the longest synset is a nominal one with 28 literals.

Table 3 contains a comparison between the level of polysemy when taking into account all literals vs. considering only polysemous ones. The comparison shows that while English literals are on average more polysemous than the French ones, there are big differences between English and French verbs, suggesting that automatically gen-

erated French verbal synsets contain some noise which will have to be filtered out in the future.

	PWN 3.0	WOLF new
avg. poly. + mono.	1.39	1.28
N	1.23	1.19
V	2.17	3.36
avg. poly. - mono	2.91	2.11
N	2.77	1.84
V	3.57	5.0

Table 3: Results of the wordnet extension procedure

A comparison of unique literals in PWN and WOLF shows that we were able to automatically generate as much as 25% of all multi-word expressions and over 30% of proper names found in PWN, which is a very good result, considering that the only source of both of these groups of literals was Wikipedia.

6.2 Manual evaluation of the results

In this section we report the results of manual evaluation of the wordnet extension where we evaluate the accuracy of the (*literal, synset*) candidates we obtained with the classifier as well as the accuracy of the candidates we discarded. For the evaluation we randomly selected 400 hundred (*literal, synset*) and evaluated them manually, using only two tags: “OK” if it would be correct to add that literal to the synset, and “NO” if it would be wrong, regardless of what the reason was for the error and how semantically close it was to the synset. The accuracy of a set of candidates is as usual as the proportion of candidates receiving the “OK” tag. Moreover, in order to assess the quality of our scoring technique, we compared the accuracy of the candidates per quartile w.r.t. their certainty scores.

The results of manual evaluation are shown in Table 4. They show a strong correlation between the certainty score they received and the accuracy of the candidates, thus justifying our decision to use this threshold but other threshold values could have been used too: higher values would have provided candidates with an even higher accuracy but the scale of the wordnet extension would have been lower; on the other hand, lower threshold values would have extended our wordnets even more, but would have introduced much more noise.

No. of candidates evaluated	400
No. of candidates added to wordnet	27%
Accuracy of all candidates	52%
Acc. of the candidates added to WOLF	81%

Accuracy of the discarded candidates	40%
Accuracy in the upper (4 th) quartile	83%
Accuracy in the third quartile	63%
Accuracy in the second quartile	41%
Accuracy in the lower (1 st) quartile	20%

Table 4: Manual evaluation of (*literal, synset*) candidates generated for extending WOLF

6.3 Automatic evaluation of the results

In this section we report the results of automatic evaluation of the generated wordnet against the already existing wordnet for French that was developed within the EuroWordNet project. With this evaluation we will gain an insight into the precision and recall of the wordnet we created with the proposed extension procedure. However, such an evaluation is only partial, because the detected discrepancies between the two resources are not only errors in our automatically created wordnets but can also stem from a missing literal in the resource we use for comparison. Automatic evaluation was performed on non-empty synsets, which means that adjectival and adverbial synsets in WOLF could not be evaluated this way at all because other existing French wordnets do not cover them.

When considering non-empty synsets in FWN, any (literal, synset) pair that is common to both resources is considered correct. When the number of valid (literal, synset) pairs of all types are combined, we reach a total of ~65,690 valid pairs out of 76,436, reaching a ~86% accuracy. A direct comparison to other related resources developed by Navigli and Ponzetto (2010) and di Melo and Weikum (2010) is not straightforward because even though the resources we used overlap to a great extent, their aim was to create a multilingual network while we focused only on French. An important difference between our approach and the one proposed by Navigli and Ponzetto (2010) is that they machine-translated the missing translations, while we only use resources that were created by humans, which is why we have more accurate translations. On the other hand, while di Melo and Weikum's (2010) wordnet for French has a slightly higher accuracy, it is smaller than ours. This shows that the approach we used to benefit as much as possible from available resources using basic NLP tools only is very efficient for building large-scale reliable wordnets.

	Correct (<i>literal, synset</i>) pairs in WOLF and FWN	Correct WOLF pairs not in FWN	Incorrect WOLF pairs not in FWN	Correct FWN pairs not WOLF
Nominal pairs not empty in FWN	8,474	11,627		15,474
		~7,441	incorrect pairs: ~4,186	
	correct pairs: ~15,915			
Verbal pairs not empty in FWN	1,826	3,859		6,168
		~1,351	incorrect pairs: ~2,508	
	correct pairs: ~3,177			
Empty pairs in FWN	0	50,650		0
		~46,598	incorrect pairs: ~4,052	
	correct pairs: ~46,598			
All pairs	10,300	66,136		21,642 + the no. of literals missing in synsets not covered by FWN
		~55,390	incorrect pairs: ~10,746	
	~65,690 <i>overall precision: ~86 %</i>			

Table 5: Automatic evaluation of the extended WOLF based on FWN

7 Results and evaluation

In this paper we described an approach to extend an existing wordnet from heterogeneous re-

sources. Using various features such as distributional similarity, we were able to reuse automatically extracted bilingual lexicons for translating and disambiguating polysemous literals, which

had so far been dealt only with word-aligned corpora. The result of our work is a freely available lexical semantic resource that is large and accurate enough for use in real HLT applications. Compared to other similar resources for French, our wordnet is bigger than the much older French WuroWordNet and more comprehensive than the much more recent JAWS database. Due to the multiple human-produced resources which it was based on is more accurate than BabelNet (Navigli and Ponzetto, 2010) and larger than the French part of the multilingual wordnet developed by di Melo and Weikum (2010).

Analysis and evaluation of the approach shows that it is both versatile and accurate enough to successfully extend a wordnet of limited coverage. Another major advantage of the approach is that it is fully modular, adaptable and language independent and can therefore be used for any language still lacking a substantial wordnet.

In the future we plans to adapt the distributional similarity measure in order to automatically detect literas that are outliers in synsets and should therefore be removed from the developed wordnet. This procedure will provide an even more accurate and useful source of the much needed lexical knowledge that is much needed in virtually all HLT tasks.

Acknowledgments

The work described in this paper has been funded in part by the French-Slovene PHC PROTEUS project “Building Slovene-French linguistic resources: parallel corpus and wordnet” (22718UC), by the French national grant EDyLex (ANR-09-CORD-008) and by the Slovene national postdoctoral grant (Z6-3668).

References

- Delphine Bernhard and Iryna Gurevych (2009). Combining Lexical Semantic Resources with Question & Answer Archives for Translation-Based Answer Finding. *Proc. ACL-IJCNLP'09*.
- Hal Daumé III (2004). *Notes on CG and LM-BFGS optimization of logistic regression*.
- Mona Diab (2004). The Feasibility of Bootstrapping an Arabic WordNet leveraging Parallel Corpora and an English WordNet. In *Proc. of NEMLAR*, Cairo, Egypt.
- Helge Dyvik (2002). Translations as semantic mirrors: from parallel corpus to wordnet. Revised version of paper presented at *ICAME 2002*, Gothenburg, Sweden.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, Massachusetts, USA.
- Adriano Ferraresi, Silvia Bernardini, Giovanni Picci and Marco Baroni (2010). Web Corpora for Biligual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation. In Xiao, R. (ed.) *Using Corpora in Contrastive and Translation Studies*. Newcastle: Cambridge Scholars Publishing.
- Darja Fišer and Sagot Benoît (2008). Combining multiple resources to build reliable wordnets. In *Proc. of TSD'08*, Brno, Czech Republic.
- Pascale Fung (1995). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proc. of ACL'95*.
- Nancy Ide, Tomaž Erjavec, Dan Tufis (2002). Sense Discrimination with Parallel Corpora. In *Proc. of the ACL '02 Workshop on Word Sense Disambiguation*, Philadelphia, USA.
- Kevin Knight and Steve K. Luk (1994). Building a large-scale knowledge base for machine translation. In *Proc. of AAAI'94*.
- Gerard de Melo and Gerhard Weikum (2009). Towards a universal wordnet by learning from combined evidence. In *Proc. of CIKM'09*.
- Mortaza Montazery and Heshaam Faili (2011). Unsupervised Learning for Persian WordNet Construction . In: *Proc. of RANLP'11*, Hissar, Bulgaria.
- Claire Mouton and Gaël de Chalendar (2010). JAWS: Just Another WordNet Subset. In *Proc. of TALN'10*, Montreal, Canada.
- Roberto Navigli, Simone Paolo Ponzetto: BabelNet: Building a Very Large Multilingual Semantic Network. In *Proc. of ACL'10*, Uppsala, Sweden.
- Simone Paolo Ponzetto and Roberto Navigli. 2009. Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proc. of IJCAI'09*.
- Maria Ruiz-Casado, Enrique Alfonseca, Pablo Castells (2005). Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. In *Proc. of AWIC'2005*. pp. 380–386
- Dan Tufis (2000): BalkaNet — Design and Development of a Multilingual Balkan WordNet. In *Romanian Journal of Information Science and Technology Special Issue*, 7(1-2).
- Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer, Dordrecht, The Netherlands.
- Dominic Widdows and Kathleen Ferraro (2008). Semantic vectors: a scalable open source package and online technology management application. In *Proc. of LREC'08*, Marrakech, Morocco.

A Novel Approach for Document Classification using Assamese WordNet

Jumi Sarmah

Dept. of Computer Science
Gauhati University
Assam; India 781014
jumis884@gmail.com

Navanath Saharia

Dept. of Computer Science
Gauhati University
Assam; India 781014
navanathsaharia@gmail.com

Shikhar Kr. Sarma

Dept. of Computer & IT
Gauhati University
Assam; India 781014
sks001@gmail.com

Abstract

The tremendous growth of the Internet led to the development of million-million of documents and so, there is a great need in developing tools that help users in searching the Web and retrieve relevant documents. Document Classification – the assignment of natural language documents to one or more pre-defined categories based on their content is an important task to gain relevant documents on the input of a query. This paper explores a novel approach for classification of Assamese documents using Assamese WordNet. The commonly used categories like *News*, *Sports*, *Science and Arts* were taken for our approach. The proposed method when applied on Assamese documents outcomes with a result of accuracy 90.27 %.

Keyword: Assamese WordNet, Document Classification.

1 Introduction

Documents are the repositories of information, and the amount of documents are rapidly and continuously increasing with the wide availability of the WWW and Internet. Document Categorization/Classification task is to assign a document to one or more predefined categories depending on its semantic contents. If the documents are properly categorized or classified then documents can be retrieved with less consuming time and with less effort. The existing literature says that the human intervention till 80's have led the researchers of 90's proposed many machine learning techniques or classifiers to automatically categorize the documents to predefined categories. The advantages of machine learning approaches are that no human intervention is required for the task of document classification.

Those systems (computers and other computing devices) that hope to process natural languages as people do must have information about words and their meanings. This information is traditionally provided through dictionaries, which are readable by machines and are widely available now. An optimal approach now can be considered to be using a WordNet which in many directions provides better facilities than a dictionary. Wordnets have been developed for various languages, and more and more Wordnets are being available with sound technicalities and architectures.

WordNet is a large lexical database viz. Database of words, which was created by professor, George A. Miller in 1985 to make a combination of dictionary and thesaurus. It groups words into a set of synonyms called *Synsets*, each shows a distinct *Concept* and then records the various semantic relation between these synsets.

For an Indian regional language Assamese, the Assamese WordNet is being developed which is a part of Indo-WordNet. Our work focuses on using Assamese Wordnet for classifying Assamese documents.

The rest of the paper discusses as follows. Section 2 portrays the Assamese WordNet. Section 3 discusses the approaches of Document Classification, Section 4 presents the system architecture of our proposed Classification approach. Experiments were done using Assamese documents in Section 5 and then results were analyzed. The paper is concluded in Section 6.

2 Assamese WordNet

The structure of Assamese WordNet was first developed by (Sarmah et al.,2010) in Gauhati

University. Assamese WordNet contains contents that are linked to the Hindi and English Wordnets. Among the major components of WordNet, some of them are described by (Hussain et al., 2011). They are:

1. ID: It is the Unique Identification Number for each word contained in the WordNet.
2. CAT: It contains the category-the Parts of Speech. The data in Assamese WordNet is separated into four parts of speech-noun, verb, adjective and adverb.
3. Synset: Synsets contains synonymous Assamese words and the words in a synset are arranged according to the frequency it is being widely used. Synsets are important in WordNet as they are considered to be the basic building blocks of WordNet. There are various lexical and semantic relations between these synsets of WordNet such as Hyponym-Hypernym, Meronym-Holonym, Entailment, Troponym, Antonym.
4. Gloss: It describes the Concepts of the given synset. Gloss plays an important role since it is through this the synsets are liked across WordNets. It contains Text Definition- It explains concepts denoted by the synset, Example Sentence - It makes use of the synsets words in the sentence.

ID: 3945

CAT: NOUN

Synset: {খারু , কংকণ,কঙ্গণ }
{Bangles}

Gloss:

Text Definition : হাতত পিঞ্চা এবিধ গহনা

A hand wearing ornament.

Example Sentence: শীলাই সোণৰ খারু পিঞ্চি আছিল

Shila was wearing golden bangles
In the above example, Synset contains the first word “খারু”(kharu: Bangles),” since it is the most frequently used word in Assamese language in the set of synset.

The above format is used to update the WordNet by the Lexicographer's with new synsets, gloss. But, Assamese WordNet is a database of words which is actually contained in three text files-Index file, Data file and Ontology. Each file have their own Individual syntax for storing words and each syntax is given a pictorial view (shown below) and marked in numerals in ascending order for which follows the other in the syntax for each file.

1. Index File: Contains details of every word in the WordNet in this file.

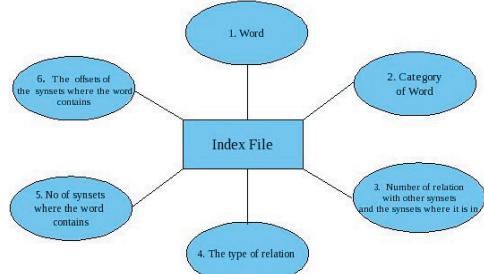


Figure 1: Index file.

2. Data File: It contains synsets and glosses with different relations. Here, this format have a symbol 'l' which means it divides gloss ,examples from others.

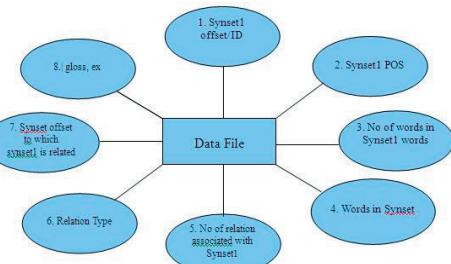


Figure 2: Data file.

3. Ontology: Ontology is a hierarchical organization of Concepts. Concepts are classified into categories, then sub-categories, and then sub-sub-categories and so on.

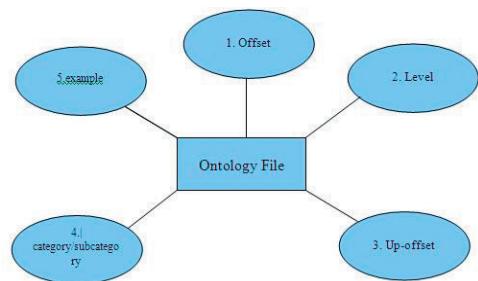


Figure 3: WordNet ontology.

Here offset means the ID of the category/subcategory; level denotes - 0000 is the top level and 0001 is the lower level. Up-offset label denotes the offset of the upper level or the super - category of it. There is the symbol 'l' in this syntax which divides the category/subcategory

and example. Each word contains a category. A category contains a subcategory and a subcategory contains a sub-subcategory so on.

The present Assamese Wordnet contains synsets of about 20,675.

3 Document Classification Approaches

Document Classification tasks can be done in two ways - Supervised document classification, where some external mechanism provides information for correct classification of documents; Unsupervised document classification, where classification is done entirely without reference to any external means. Also a new approach called Semi-Supervised document classification has been sorted where the learning scheme lies somewhere between supervised and unsupervised approach. Normally, Document Classification is done by a supervised way where predefined categories are taken for implementation. During the Literary Survey we came across two most widely used approaches for classifying documents - 1. Machine Learning Approach and 2. WordNet based Approach.

3.1 Machine Learning Approach

One of the branch of Artificial Intelligence is Machine Learning, a scientific discipline concerned mainly with developing algorithms so that computers get adapted with some behavior. Many machine learning techniques for classifying documents have been proposed. Some of them are – Neural Network (Sebastiani, 2002); Self-Organizing Map (SOM) is one of the widely used Neural Network algorithms (ChandraShekar et al., 2009), Decision Tree (Sebastiani, 2002; Baharudin et al., 2010), Naive Bayes Classifier(Hotho et al., 2005; Rigutini, 2004), Nearest Neighbor Classifier (Sebastiani, 2002; Rigutini, 2004), Centroid Based Classifier (Sebastiani, 2002; Joachims, 1998), Support Vector Machine (Rigutini, 2004; Joachims, 1998).

For Machine Learning approach, the authors of paper(Harish et al., 2010) concludes saying that SVM(Support Vector Machine) outperforms well in many Document Classification tasks.

3.2 WordNet based approach

Document Classification using the approach of WordNet has been performed by the authors of (Elberrichi et al., 2008; Mohanty et al., 2005; Peng et al., 2005). A approach using the

Hypernyms of WordNet is performed by(Scott and Matwin, 1998), a hybrid technique -KNN and Wordet Senses, by (Rosso et al., 2004) and found an accuracy of 82%. Also SOM and WordNet by (Brezeale, 1999) were used to categorize documents to some predefined categories. WordNet was also used to disambiguate the word meanings/senses for improving categorization of documents by (Liul et al., 2007). The author's of the paper (Li et al., 2009) tried to automatically classify documents and found an accuracy of more than 90%.

4 System Architecture and Design

Our proposed architecture can be divided into three phases-

1. Pre-processing phase
2. Tuned Categories
3. Classification phase

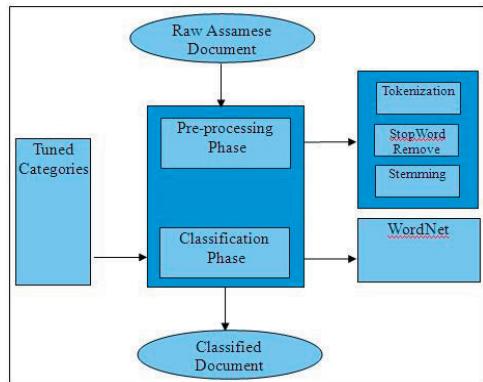


Figure 4: Architecture of our system.

The design phases of our proposed architecture is discussed below:

4.1 Pre-processing of Document Classification

Before classifying a document we should follow these following pre-processing steps to make them machine comprehensible.

Tokenization: Tokens are the units of meaningful letters and tokenization is the process which helps to separate each word from the other words in a document. When tokenized it helps to do other pre-processing steps fluently. But, some words like -"লাহে-লাহে"(lahe-lahe: slowly) when applied tokenization do not tokenize such words but take as one word – which is a problem yet to be overcome.

Stop Words Removal: Stop Words are those words that do not play an important role in retrieving information. Stop Words in Assamese language are like “আৰু” (*aaru* : and), “ তথা (*to-tha*:Hence)”, “হয় (*hoi*:Yes)”, etc.

Sorting: Sorting the terms in ascending order after stop word removal helps in doing stemming of words.

Stemming: It reduces the tokens to their root form so that it helps in counting frequency of the terms together having the same word sense . A partial stemmer for Assamese is developed for our work. At first we have first checked each word if it ends with Assamese suffixes like “খিনি”, “ৰ”, “বোৰ”, “বিলাক”, “লোক”, “সকল” etc. then, if so, we removed those suffixes and then go on checking the sorted words that are below the emitted suffix word. Then, if the below words begins with the emitted-suffix word then we count the frequency along with the “emitted-suffix word” as those words have same sense (meaning). The word not beginning with the “emitted suffix word” are repeated once again by removing the suffixes and so on. A thing to remember here is that a stemmer never stems a word to its correct meaning. If it does so then it is called a Lemmatiser. But, this stemming algorithm developed for our approach does not stem words containing conjunctive suffixes.

Threshold value: The most frequent terms to a certain threshold value are taken. The threshold value here is the total number of words in the document.

4.2 Tuned Categories

As the document classification is normally done in supervised way, the predefined categories taken for classification task should be tuned at first so that document could be classified to one of those predefined tuned-categories. At first, each categories should pass through preprocessing phase. Then, the terms/words in the categories should be checked if it exists in Assamese WordNet's Synsets. Assamese WordNet is divided into four categories Core Word Net, Common,PAN-Indian and Universal. Each term should be checked in these categories of WordNets except Core because all words en-tried in Core are also in Universal. Thus, the tuned-categories-News, Sports, Science and Arts(taken for our approach) are built consisting of those terms that are in Assamese WordNet. The tuned categories have been annotated with the below

number of synsets- Arts-163, News-138, Sports-136, Science-153. The tuned category 'News' contains those news that are related to 'political affairs'.

4.3 Classification Phase

In Document Classification phase the input document to be classified to a particular tuned category should also go through the pre-processing steps and then check out for its frequent terms if it exists in the synsets of Assamese Wordnet and then for a particular frequent term t_i if found in WordNet's synset, creates an extended form of the term(t_i) to ext_i , (term and its synsets where it occurs). The extended term when found are then tokenized so that each term could be checked to the predefined tuned-categories. The highest number of matching terms in a particular tuned-category is the relevant category where the document should be categorized to.

Below an example explains the phase-

Say the frequent term “খেল” found in the WordNet's synset for the document to be classified after pre-processing phase. Then, it is merged with the synset to form ext_i that contains “খেল, ৢৰীড়া, খেল-ধেমালি তামাচা, খেল খেল, ৢৰীড়া, খেল-ধেমালি”. Then each word tokenized from ext_i , checked with the frequent WordNet mapping terms of the tuned categories. The category that outcomes with the highest number of matching terms is the category for the document.

5 Results and Analysis

We prepared our tuned model corpus from two daily Assamese news papers, viz.- “*'Assamiya_Pratidin'* and ‘*Khabar*’ ”. The input documents are taken from various sources like Assamese Blogs, the Sunday Indian magazine and tested. We have tested by taking 20 files each with separate domain and tried to categorize it to some common categories like Sports, News, Science and Arts. Experimental Results are determined with “ F_1 measure”. F_1 measure is the harmonic mean of Precision and Recall where Precision and Recall are two widely used measures used in Document categorization literature to evaluate the algorithm's effectiveness on a given category and it is referenced in paper(Elberrichi et al., 2008). Precision and Recall can be formulated as:

$$F1 \text{ (Recall , Precision)} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Precision=(true positive/true positive+false positive)

Recall= (true positive/true positive+false negative)

where, true positive- the number of items correctly labeled as belonging to positive class, false positive-are the items incorrectly labeled as belonging to other class, false negative-the items which were not labeled as belonging to the positive class but it should have been belonged to.

F_1 measure is used to evaluate the overall performance of our approach. According to existing literature, macroaverage F_1 computes the F_1 value for each category and then take average of the per category score of F_1 . Given a Corpus with n categories, we let the F_1 value for the i-th category be $F_1(i)$, macroaverage $F_1(i)$ is defined as-

$$\text{macroaverage } F_1(i) = \sum_{i=1}^n F_1(i) / n$$

Here we used the most common categories of the Corpus for training and testing.

For calculating the F_1 measure we used the the following table information. The first column indicates the serial No. of the tested files, second column indicates the total number of words the file belongs, the third column indicates the resultant domain and in the fourth column we tag with Correct/Incorrect label which indicates the correctness/in-correctness of our classification task's output. In the category column the category written in brackets should be the actual category for the document.

Sl. No.	No. of Words	Category	Result
1	742	Sports	Correct
2	1088	News	Correct
3	814	News	Correct
4	1419	Arts	Correct
5	1478	Arts	Correct
6	789	Science	Correct
7	616	Science	Correct
8	702	Arts	Correct
9	1047	Science(Arts)	Incorrect
10	843	News	Correct
11	1248	Sports	Correct
12	948	Sports	Correct

13	1001	Science (Sports)	Incorrect
14	1225	Sports	Correct
15	1759	News	Correct
16	942	News	Correct
17	1841	Arts	Correct
18	2332	Science	Correct
19	2483	Science	Correct
20	703	Science	Correct

Table 1: Table showing results of 20 tested files.

Cate-gories	Precision	Recall	F_1 measure
Arts	100%	80%	88.88%
Sports	100%	80%	88.88%
Science	71.42%	100%	83.33%
News	100%	100%	100%

Table 2: Table showing the F_1 -measure for the four categories.

Macro-average $F_1 = 90.27\%$.

6 Conclusion and future works

Document Classification using WordNet is a significant step to the Assamese Language because till date no task for classifying document in Assamese language has been done. Here, we have categorize twenty input documents to pre-defined categories. But, we can also sub-categorize a category, say Sports can be further categorized to Cricket if it contains a lot of information on the International play Cricket, also plays like Hockey etc. By developing a such application we can further increase the performance of document classification task. The Assamese WordNet is in developing stage. When all the Assamese unique words and their synsets will be covered by the Assamese WordNet then our classification task's accuracy could be significantly improved. The predefined categories should contain texts of large number of words, If the categories size could be increased, this would be another step to improve the performance of our Document classification task. Using the advantage of WordNet we tried a new approach - To classify document to its relevant category and

through which we were successful to a reasonable amount.

Acknowledgement: The authors acknowledges the TDIL programme of Department of Information Technology, Ministry of Communication and IT of Government of India for funding the Wordnet Project on Assamese Wordnet Development currently being executed at Gauhati University.

References

- Andreas Hotho, Andreas Nurnberger and Gerhard Paaf. 2005. *A Brief Survey of Text Mining*. Journal of Machine Learning (2005), 20(1):19-62 Publisher: s.n., ISSN: 01751336.
- B. H. ChandraShekar and Dr.G. Shoba. 2009. *Classification Of Documents Using Kohonen's Self-Organizing Map*. International Journal of Computer Theory and Engineering, Vol.1, No.5, 610-613.
- B. S. Harish, D. S. Guru, S. Manjunath. 2010. *Representation and Classification of Text Documents:A brief Review*. International Journal of Computer Application, RTIPPR(2): 110-119.
- Baharum Baharudin, Lam Hong Lee and Khairullah Khan. 2010. *A Review of Machine Learning Algorithms for Text-Documents Classifications*. Journal of Advances in Information Technology, Vol. 1, No. 1, 4-20.
- Darin Brezeale. 1999. *The Organization of Internet Web Pages using WordNet and self organizing Maps*. Computer Science and Engineering, University of Texas at Arlington, 1999.
- Fabrizio Sebastiani. 2002. *Machine Learning In Automated Text Categorization*. Journal of ACM Computing Surveys, Vol.3, No.1, 1-47.
- Iftikaar Hussain, Navanath Saharia and Utpal Sharma. 2011. *Development of Assamese WordNet*. Machine Intelligence:Recent Advances, Narosa Publishing House, Editors. B. Nath, U. Sharma and D. K. Bhattacharyya, ISBN-978-81-8487-140-1, 2011.
- Jianqiang Li, Zhao Yu and Bo Liu.2009.*Fully Automatic Text Categorization by Exploiting WordNet*.
- 5th Asia Information Retrieval Symposium, (AIRS 2009), 1-12.
- Leanardo Rigutini.2004. *Automatic Text Processing Machine Learning Techniques*. Information Engineering at Department of Information Engineering of the University of Siena (Italy).
- Paolo Rosso, Edgardo Ferretti, Daniel Jimenez and Vicente Vidal. 2004. *Text Categorization and Information Retreival using WordNet Senses*. 2nd Global Wordnet Conference (GWC'04), Vol. 2945, 299-304.
- S. Mohanty, P. K. Santi, Rajneeta Mishra, N.R. Mohapatra and Sabyasachi Swain. 2005.*Semantic Based Text Classification Using WordNets : Indian Language Perspective*. Global WordNet Conference(GWC2006), 321-324.
- Sam Scott, and Stan Matwin.1998.*Text Classification using WordNet Hypernyms*. Association for Computational Linguistics (ACL 1998),45-52.
- Shikhar Kr Sarma, Moromi Gogoi, Utpal Saikia and Rakesh Medhi, 2010. *Foundation and structure of Developing Assamese WordNet*. 5th International Conference of the Global WordNet Association (GWC-2010).
- Songbo Tan. 2007. *An Improved Centroid Classifier for Text Categorization*. Expert System with Application(Expert Syst. Appl.), 279-285.
- Thorsten Joachims.,1998. *Text Categorization with Support Vector Machines:Learning with many relevant features*. The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases – ECML.
- Xiaogang Peng and Ben Hoi. 2005. *Document Classification based on Word Semantic Hierarchies*. International Conference on Artificial Intelligence and Applications(IASTED), 362-367.
- Ying Liul, Peter Scheuermann, Xingsen Li and Xingquan Zhu. 2007. *Using WordNet to Disambiguate Word Senses for Text Classification*. International Conference on Computational Science, (ICCS 2007), 781-789.
- Zakaria Elberrichi, Abdelattif Rahmoun and Mohamed Amine Bentalah. 2008. *Using WordNet for Text Categorization*. The International Arab Journal of Information Technology, Vol. 5, No. 1, 16-24.

Refining WordNet adjective dumbbells using intensity relations.

Vera Sheinman and Takenobu Tokunaga

vera@sheinman.org, take@cl.cs.titech.ac.jp

Tokyo Institute of Technology, Tokyo, 152-8552 Japan

Isaac Julien and Peter Schulam and Christiane Fellbaum

ijulien, pschulam, fellbaum@princeton.edu

Computer Science Department, Princeton University, New Jersey 08540, USA

Abstract

We propose a new semantic relation of intensity for gradable adjectives in WordNet and show its specific benefits for NLP. Intensity enriches the present, vague, *similar* relation with information on the degree with which different adjectives express a shared attribute. Using lexical-semantic patterns, we mine the Web for evidence of the relative strength of adjectives like “large”, “huge” and “gigantic” with respect to their attribute (“size”). The pairwise orderings we derive allow us to construct scales on which the adjectives are located.

To represent the intensity relation among gradable adjectives in WordNet, we combine ordered scales with the current WordNet *dumbbells* based on the relation between a pair of central adjectives and a group of undifferentiated semantically *similar* adjectives. A new intensity relation links the adjectives in the dumbbells and their concurrent representation on the scale.

1 Introduction

A survey of publications on NLP work using WordNet shows that the more than 18,000 adjective synsets are rarely part of a system, and many crosslingual wordnets do not include adjectives at all. This may be attributable to the role of adjectives as modifiers and carriers of arguably less essential information. But we conjecture that one principal reason for the current under-use is that the organization of adjectives in WordNet does not lend itself well to a clear determination of semantic similarity. The present work explores the semantics of scalar adjectives and outlines a novel way of representing such meanings in WordNet.

1.1 Adjectives in WordNet

WordNet originated as a model of human semantic memory. Specifically, it was designed to test then-current models of conceptual organization that supported a network structure (Collins and Quillian, 1969). Association data indicated that words expressing semantically similar concepts were stored in close proximity and strongly evoked one another. Thus, when presented with a stimulus word like “automobile”, people overwhelmingly respond with “car”; the prevalent response to “celery” is “vegetable” and to “elephant”, “trunk” (Moss and Older, 1996). Such data suggested the organization of words and concepts into a network structured around semantic relations like synonymy, meronymy (part–whole) and hyponymy (super/subordinates).

Most striking is the strong mutual association between members of antonymous adjective pairs like “wet–dry” and “dark–light”, already discussed by (Deese, 1964) who noted that such pairs are acquired early by children. The *clang* association between antonymous adjectives might well be due to their high frequency and their shared contexts that indicate their common selectional restrictions. (Justeson and Katz, 1991) showed furthermore that members of an antonymous adjective pair co-occur in the same sentence far more often than chance would predict.

It seemed straightforward enough to represent the members of an antonym pair as opposite poles on an open-ended scale that encoded a particular attribute. But what about the many adjectives that are semantically similar to these adjectives yet are neither synonyms nor antonyms of a member of the pair?

(Gross et al., 1989) measured the time it took speakers to respond to questions like “Is small the opposite of large?”, “Is miniature the opposite of large?” and “Is gigantic the opposite of

miniature?” The first kind of question involved the members of an antonym pair and the latencies here were very short. The second kind of question involved one member of an antonym pair and an adjective that was similar to its antonym. People took measurably longer to affirm these questions. The third kind of question asked people’s judgments about two adjectives that were each similar to one member of an antonym pair. In these cases, people either were hesitant to reply at all or they took a very long time to respond affirmatively.

These data inspired the representation of adjectives in WordNet by means of *dumbbells*, with antonyms as the centroids and semantically similar adjectives arranged in radial fashion around each antonym. Figure 1 depicts a schematic representation of a dumbbell.

1.2 Limitations of the Dumbbell Representation

While the dumbbells seemed well motivated psycholinguistically and distributionally, they do not lend themselves easily to Natural Language Processing and they stump systems designed to detect and quantify meaning similarity.

First, relatively few adjectives are interconnected, which limits path-based Word Sense Disambiguation systems to the small number of adjectives that are classified as being either antonyms or semantically similar in a given cluster. Second, within a cluster, all semantically similar adjectives are arranged equidistantly from a centroid. As a result, the path length between the centroid and all similar adjectives is always one and that between two similar adjectives is invariably two, with each path connected via the centroid. This lack of encoding of independent meaning distinctions among the *similar* adjectives suggests that they are all equally similar to the centroid, which is intuitively not the case. For example, both “titanic” and “capacious” are represented as being equally similar to “large”, as are “subatomic” and “gnomish” to “small”. Moreover, the meaning differences among the similars themselves, such as “titanic”, “capacious”, “monstrous” and “gigantic” on the one hand, and “subatomic”, “gnomish”, “dinky” and “pocket-size” on the other hand, are not represented. Finally, many similar adjectives are in fact misclassified as members of a same cluster, whereas based on their selectional restrictions, they should in many cases be assigned to

different clusters. Thus, “hulking” describes entities with physical properties, while a related similar adjective like “epic” typically modifies abstract concepts like events (“epic battle”, “epic voyage”). Likewise, adjectives that are currently classified as being similar to “small”, for example “pocket-size” and “elfin”, differ in their selectional restrictions: the former can be applied to objects like books, whereas the latter typically modifies people.

Semantically, the relation of the centroids to the similar adjectives as well as that among the similar adjectives themselves is underspecified and expressed only indirectly via antonymy. A second relation, labeled *see also* links different dumbbells via a shared centroid adjective that has a different but related sense in each dumbbell. It is often difficult to discern a motivated distinction between the similar and the *see also* relations and hence, among the adjectives they connect.

1.3 Scalar Adjectives

Our focus here is on adjectives that possess scalar properties. (Bierwisch, 1989) notes that dimensional adjectives like “long”, “short”, “wide”, “narrow”, “new” and “old” express a particular value on a scale or dimension. For example, while both “ancient” and “old” fall on the same scale (“age”), their relative placement on the scale represents the fact that “ancient” expresses a more intense degree of “age” and “old”.

Some dimensional scales lexicalize many points (“large-small”), while others express few points besides paired polar antonyms (“tall–short”). Note that the scales are open-ended, and a stronger or weaker degree of the underlying shared attribute can always be conceived of, even if it is not independently lexicalized.

We propose a re-organization of the subset of adjectives that express different values of a gradable property (Bierwisch, 1989; Kennedy, 2001) using the AdjScales method (Sheinman and Tokunaga, 2009). For a given attribute, we construct scales of adjectives ordered according to the intensity with which they encode a shared attribute. The ordering will be based on corpus data.

2 AdjScales

The AdjScales method orders a set of related adjectives on a single scale using the intensity relation, as in the example *tiny* → *small* →

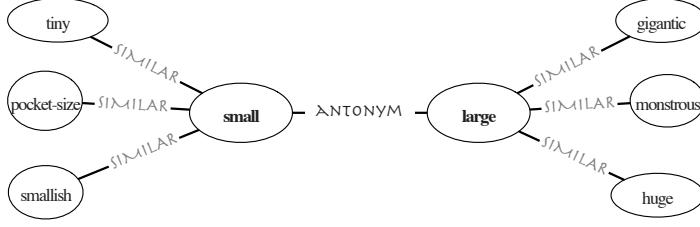


Figure 1: An illustration of WordNet’s dumbbell structure.

smallish → *large* → *huge* → *gigantic*.

The basic methodology of AdjScales is to extract patterns characterizing semantic relations from free text based on several word instances, and then use the extracted patterns for extraction of further instances of the relations of interest, or even for bootstrapping of additional patterns. Several techniques for extracting semantic similarity from corpora have been proposed.

Lexical-semantic patterns were first described by (Cruse, 1986), who notes that phrases like “*xs* such as *ys*” and “*ys* and other *xs*” identify *x* as a superordinate, or hypernym, of *y*. (Hearst, 1992) pioneered the identification and application of such phrases or patterns to the extraction of semantically related words from corpora as an efficient way to semi-automatically construct or enrich thesauri and ontologies. Her work was further extended by (Riloff and Jones, 1999; Chklovski and Pantel, 2004; Turney, 2008; Davidov and Rappoport, 2008; Snow et al., 2005).

Contextual or distributional similarity based approaches such as (Weeds and Weir, 2005; Lin, 1998) rely on the observation that words with similar meanings also share similar contexts. For instance, “*rose*” and “*flower*” constitute a hyponym-hypernym pair and thus one can expect some of the contexts of “*rose*” to appear in the same contexts of “*flower*”; differently put, semantically similar words are often mutually interchangeable.

Pattern-based extraction method identifies words that are *paradigmatically* related; approaches based on contextual similarity rely on the *syntagmatic* similarity of related words.

Both approaches to identifying semantically similar words should converge; automatically derived thesauri such as (Lin, 1998) show significant overlap with manual resources like WordNet. The AdjScales method exemplifies the phrase-based extraction approach.

AdjScales comprises two stages, preprocessing and scaling that are described in detail in (Sheinman and Tokunaga, 2009). The following section summarizes the method with an eye towards enriching adjectives in WordNet with intensity information.

2.1 Preprocessing: pattern extraction

The preprocessing step of the AdjScales handles extraction of patterns that later serve AdjScales for scaling of adjectives. Pattern extraction queries of the form “*a * b*” are used, where *a* and *b* are seed words and “*” denotes a wildcard (zero to several words that may appear in its place). AdjScales extracts binary patterns of the form

$$p = [\text{prefix}_p \quad x \quad \text{infix}_p \quad y \quad \text{postfix}_p]$$

from the snippets of the query results using a search engine, where *x* and *y* are slots for words or multiword expressions. A pattern *p* can be instantiated by a pair of words *w*₁, *w*₂ to result in a phrase “*prefix*_{*p*} *w*₁ *infix*_{*p*} *w*₂ *postfix*_{*p*}”.

Let us consider an example pattern *p*₁ where prefix_{*p*₁} = ϕ , infix_{*p*₁} = “if not”, and postfix_{*p*₁} = ϕ , if we instantiate it with the pair of words (large, gigantic) we will get a phrase *p*₁(large, gigantic) = “large if not gigantic”.

If *p(w*₁, *w*₂) appears in snippets that are returned by a search engine when querying it with a pattern-extraction-query, we refer to it as *p* is supported-by (*w*₁, *w*₂). For the extraction purposes snippets are split into sentences and are cleaned from all kinds of punctuation. Up to this point, the notation and the method largely follow the work by (Davidov and Rappoport, 2008).

Differently from (Davidov and Rappoport, 2008) the seed word pairs for AdjScales are chosen in a supervised manner, so that *seed*₂ is more intense than *seed*₁. Consider, for instance the

pair (“cold”, “frigid”), where “frigid” is more intense than “cold”. The relation *more-intense-than* is asymmetric. Therefore, AdjScales selects only the *asymmetric patterns* that are extracted consistently so that the less intense word in each supporting pair is only on the left side of the pattern (before the infix words) or so that the less intense word is only on the right side of the pattern (after the infix words). Unless all the supporting pairs of words share the same direction, the pattern is discarded. The former selected patterns are defined as *intense*, and the latter as *mild*.

AdjScales selects only the patterns supported by at least 3 seed pairs and requires a pattern instance by each supporting pair to repeat at least twice in the sentences extracted from the snippets to increase reliability. It also requires the patterns to be supported by adjectives describing different attributes (seed pairs should be selected accordingly). This constraint is important, because patterns that are supported by seeds that share the same attribute tend to appear in very specific contexts and are not useful for other attributes. For instance, [*x even y amount*] might be extracted while supported only by seeds sharing the “size” attribute, such as (“huge”, “astronomical”), (“large”, “huge”), (“tiny”, “infinitesimal”).

(Sheinman and Tokunaga, 2009) report on 16 English patterns that were extracted using this stage of the method. For the analysis of the English examples presented in this work, we did not reproduce the preprocessing stage, but used the 16 patterns reported in their work and augmented them with a set of 17 human constructed patterns. Table 1 lists all the patterns used in this work.

2.2 Scaling

For this step, we use AdjScales to process the dumbbell structure from WordNet to enrich it with intensity information. We process each one of the antonymous groups in the dumbbell separately. For each pair (head-word, similar-adjective), we instantiate each pattern p in patterns that were extracted in the preprocessing stage to obtain phrases $s_1 = p(\text{head-word}, \text{similar-word})$ and $s_2 = p(\text{similar-word}, \text{head-word})$. We send s_1 and s_2 to a search engine as two separate queries and check whether $df^1(s_1) > weight \times df(s_2)$ and whether $df(s_1) > threshold$. The higher the

Table 1: Intense and mild patterns. x and y represent adjectives so that x is more intense than y .

Intense Patterns
(is / are) x but not y
(is / are) very $x y$
extremely $x y$
not x (hardly / barely / let alone) y
x (but / yet / though) never y
x (but / yet / though) hardly y
x (even / perhaps) y
x (perhaps / and) even y
x (almost / no / if not / sometimes) y
Mild Patterns
if not y at least x
not y but x enough
not y (just / merely / only) x
not y not even x
not y but still very x
though not y (at least) x
y (very / unbelievably) x

values for the *threshold*² and *weight*³ parameters, the more reliable are the results. If p is of the type *intense*, then a positive value is added to the similar-word’s score, otherwise if p is of the type *mild* a negative value is added. When all the patterns are tested, similar-words with positive values are classified as intense, while the similar-words with negative values are classified as mild. Words that score 0 are classified as *unconfirmed*. For each pair of words in each one of the subsets (mild and intense), the same procedure is repeated, creating further subsets of *mildest* words that have the most negative values within the mild subset, and *most intense* words for the words with the highest positive values within the intense subset.

After the two parts of the dumbbell are processed, they are unified into a single scale. The unification attempts to order the adjectives from the half of the dumbbell with the less frequent centroid (starting from the most intense to the mildest) to the more frequent side (starting from the mildest to the most intense). Adjectives of similar intensity are grouped together.

The adjectives in a final scale are then linked

²*threshold* regulates the number of pages returned by the search engine that is considered sufficient to trust the result, and it was set to 20 in this work.

³*weight* regulates the gap between s_1 over s_2 that is required to prefer one over the other, and it was set to 15 in this work.

¹df represents document frequency.

from the original adjective synsets in a dumbbell as illustrated in Figure 2. The unconfirmed adjectives on both sides of the dumbbell remain unlinked to the final scale.

Examples of scales extracted by applying AdjScales to the dumbbells in WordNet include:

- destitute → poor → broke → rich → loaded
- ice-cold → cold → chilly → tepid → warm → hot → (torrid, scorching)
- filthy → dirty → dingy → clean → spotless

2.3 Using the Web as a corpus

AdjScales is designed to extract fine-grained distinctions, and the relative sparseness of the lexical-semantic patterns with many of the less frequent adjectives mandates the use of a very large corpus. Second, the method requires a large, domain-independent corpus that reflects current language use and accommodates ever-shifting changes in meaning across diverse speaker communities. In particular, words with a strong flavoring tend to acquire a weaker connotation and reduced intensity with frequent use. While the Web has sometimes been criticized for being unreliable and unstable (Kilgarriff, 2007), it is a logical choice for our work, as corpora constructed for research purposes tend to be small (MASC), unbalanced (PropBank), and not representative of current language use (Brown Corpus, BNC). Finally, the method relies on the availability of a search engine that supports proximity search, provides an estimated number of page hits and snippets of the relevant Web pages.

3 Related Work

VerbOcean VerbOcean (Chklovski and Pantel, 2004) is a pattern-based approach to extracting fine-grained semantic relations among verbs from the Web. In contrast to other approaches, the patterns in VerbOcean are manually grammatically enhanced to be selective for verbs (see also (Fellbaum, 2002)). VerbOcean accounts for the frequency of the verbs as well as the frequency of the patterns themselves. Furthermore, VerbOcean distinguishes between symmetric and asymmetric semantic relations and utilizes this distinction. VerbOcean identifies six semantic relations among verbs, including *strength*, a subtype of *similarity*.

Strength, which is similar to *intensity* among adjectives, relates verb pairs in which one member

denotes a more intense, thorough, comprehensive or absolute action than the other member, as in the case of “startle” and “shock”.

A total of eight patterns were selected for extraction of the *strength* relation, including the patterns [*x even y*] and [*not just xed but yed*]. In the evaluation, the authors report that out of 14 sample pairs classified by VerbOcean as related by strength 75% were correctly classified.

Near Synonyms Differentiating between adjectives by their position on an intensity scale may fall into the research area of differentiation among *near-synonyms*. According to (Edmonds, 1999) near-synonyms are words that are alike in essential, language-neutral meaning (denotation), but possibly different in terms of only peripheral traits, whatever these may be. It is an open question whether true synonyms exist at all; WordNet defines membership in a synset as the property of being exchangeable in many, but not all contexts.

(Edmonds, 1999) introduces an extensive model to account for the differences among near-synonyms, classifying the distinctions into *denotational*, *expressive*, *stylistic*, and *collocational*. Thus, stylistic distinctions include differences in *formality*⁴. For example, “motion picture” is a more formal expression than “movie” which in turn is more formal the “flick”.

The AdjScales method indirectly takes into consideration some of the criteria for synonymy in (Edmonds, 1999). The nature of the lexical-semantic patterns is such that they retrieve snippets in which an adjective pair necessarily modifies the same noun; the narrow context moreover assures stylistic homogeneity of the scalemates.

4 Limitations of the AdjScales method

The AdjScales method promises to grant insight into a relatively underexplored corner of the lexicon by providing empirical evidence for subtle intuitions about the intensity of gradable adjectives. Scales constructed on corpus data may reflect the lexical organization of a broad community of language users. At the same time, the distinctions among the adjectives on a given scale can be very fine-grained, and speakers’ explicit judgments do

⁴WordNet’s *domain* labels encode some register and usage distinctions, but the categories are notoriously fuzzy. (Maks and Vossen, 2010) talk in detail about the differences between synset members in WordNet and propose remodeling solutions to overcome this problem.

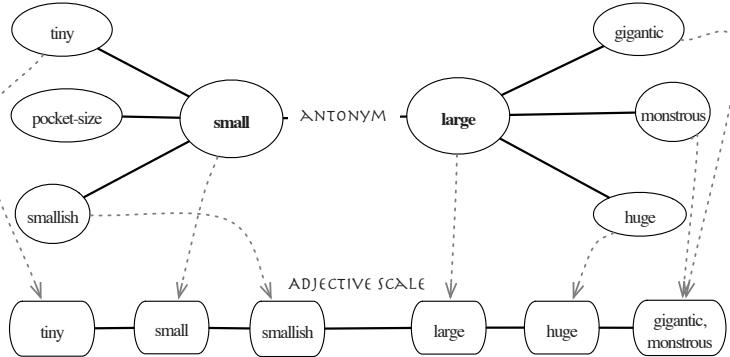


Figure 2: Illustration of the proposed structure of an adjective scale linked from some adjectives in a dumbbell. Note that “pocket-size” has more specific selectional restrictions than the other, more generically applicable adjectives in the dumbbell. It remains unconfirmed and not linked to the scale. “Smallish” is determined to be less intense than the centroid “small”. “Gigantic” and “monstrous” are recognized to be of similar intensity relatively to “huge” and “large”.

not always conform to the scales constructed on the basis of the corpus data. In the evaluation reported by (Sheinman and Tokunaga, 2009) annotators agreed with each other for only 63.5% of the adjective pairs when judging whether an adjective is milder, similar in intensity, or more intense than another adjective.

AdjScales method in particular, and pattern based methods in general, may suffer from low coverage. (Sheinman and Tokunaga, 2009) report that out of total of 5,378 distinct descriptive adjectives, only 763 were selected as suitable for further scaling, because the remainder could not be extracted in sufficient numbers in the patterns produced by the AdjScales’ preprocessing stage, which requires at least 3 seed pairs. This limitation calls for further refinement of the method, such as the extraction of a wider selection of patterns.

Another weakness of the method is its poor ability to determine the place of adjectives in the neutral area of an adjective scale. For example, “tepid”, “smallish”, and “acceptable” are difficult to properly locate on their corresponding scales, and the weakness of method here is reflected in lower human agreement. Extending our work to a larger number of attributes will show whether this problem is specific to the limited number of scales tested or more general.

Currently we apply the AdjScales method on each half of a dumbbell and unify the results into a single scale. This approach relies on the assumption that each dumbbell can produce a single scale, which is not necessarily the case. The reason is that in many cases, WordNet currently subsumes

semantically heterogeneous adjectives in a single dumbbell. Consider the adjectives “chilly, frosty, cutting, unheated” and “raw”, which are all part of a dumbbell centered around (one sense of) “cold”. But due to their different selectional restrictions, the Web does not return snippets like “* he ate his food unheated but not arctic” and “* a cutting, even refrigerated wind”. We plan to examine the members of dumbbells for their semantic similarity and refine the clusters such that they lend themselves better to placement on scales. The AdjScales method will help in the identification of semantically homogeneous adjectives, leading to a cleaner representation in WordNet.

5 Applications of AdjScales in WordNet

We discuss a representative sample of applying AdjScales to gradable adjectives below.

5.1 Language pedagogy

Adjective scales in WordNet will provide learners of English with a more subtle understanding of the meanings of adjectives. By contrast, WordNet’s current dumbbell representation and standard thesauri do not give clear information about the meaning distinctions among similar adjectives. We plan to develop a new interface that lets users visualize the unidimensional scales and gain an intuitive access to the meanings with a single glance.

5.2 Crosslingual encoding

Constructing and encoding scales with gradable adjectives for languages that have this lexical category would allow one to compare crosslinguis-

tic lexicalizations with respect to questions like: which languages populate a given scale more or less richly? How do the members of corresponding scales line up? Mapping scales across languages could well support fine-grained human and machine translation.

(Schulam and Fellbaum, 2010) extracted patterns from the large COSMAS-II⁵ German corpus using the process described in Section 2.1.

5.3 Reading textual entailment

Modeling the understanding of implicit and entailed information is a major focus of current research in NLP. The PASCAL Recognizing Textual Entailment task challenges automatic systems to evaluate the truth or falsity of a statement (the Hypothesis) given a prior statement (the Text). For example, a system must decide whether or not H is true or false given T:

- T: Frigid weather sweeps across New Jersey
- H: The Garden State experiences cold temperatures

(Clark et al., 2007; Clark et al., 2008; Fellbaum et al., 2008) show that the semantic knowledge encoded in WordNet can be harnessed to extract information that is not present on the surface. Thus, WordNet tells that “New Jersey” and “the Garden State” are synonymous, increasing the probability that H is true. But knowing that “frigid” unilaterally entails “cold” would allow a more confident evaluation of H. If T and H were switched, the symmetric synonymy relation between the nouns would not facilitate a correct evaluation of H, whereas the downward entailing intensity relation would evaluate a Hypothesis containing “frigid” to be false if the Text referred to “cold”. An RTE system with access to a resource that encodes intensity relations among its adjectives is thus potentially more powerful.⁶

⁵<http://www.ids-mannheim.de/cosmas2>

⁶Currently, WordNet encodes entailment relations among some verbs, but it doesn’t provide a distinction between finer-grained subtypes such as *backward presupposition* (“know” must happen before “forget”) vs. *temporal inclusion* (“step” is part of the action of “walk”) (Fellbaum et al., 1993). Extracting instances of specific fine-grained relations, including intensity (may → should → must) using computational methods such as those in VerbOcean may provide further enrichment of WordNet.

5.4 Identifying spam product reviews

(Julien, 2010) examines how AdjScales might be used as a tool for detecting spam product reviews. Spam reviews are online reviews of products written for either deceptive or unhelpful purposes. For instance, company owners or employees may write a positive review of their product to boost the chances that customers will buy it; conversely, negative review of a competitor’s product to discourage sales. Such reviews are more likely than genuine ones to contain highly intense adjectives.

5.5 Comparing nouns with AdjScales

(Schulam, 2011) develops a prototype of a system called SCLE (Semantic Comparison of Linguistic Entities), which uses the AdjScales algorithm to build adjective scales to compare the values represented by nouns modified by scalar adjectives.

Consider the phrases “warm day” and “hot day.” Without knowledge of the relative intensity of adjectives that ascribe different values of “temperature” to the nouns, a system may know only that both nouns are modified by semantically similar adjectives. SCLE accesses adjective scales to infer which of the two days is characterized by a higher “temperature”.

6 Conclusion

We propose a new semantic relation for WordNet’s currently under-used adjective component. The *intensity* relation holds among gradable adjectives that fall on different points along a scale or dimension. Identifying and encoding this relation relies crucially on AdjScales (Sheinman and Tokunaga, 2009), a method for extracting and applying lexical-semantic patterns to a corpus. The patterns differentiate semantically similar adjectives in terms of the intensity with which they express a shared attribute and make it possible to construct scales where the adjectives are ordered relative to one another based on their intensity.

While only gradable adjectives express varying degrees of intensity, they constitute a highly frequent and polysemous subset of adjectives that are richly encoded crosslinguistically. We propose a model for representing scales in WordNet such that they supplement and co-exist with the current dumbbells. The principal improvement will be an empirically supported refinement of the present vague *similar* relation among many adjectives arranged around a shared centroid. The en-

coding of fine-grained intensity relations among presently undifferented adjectives will greatly enhance WordNet’s potential for a wide range of diverse applications.

Acknowledgments

Fellbaum, Schulam and Julien’s work is supported by grant CNS 0855157 from the U.S. National Science Foundation; Fellbaum is additionally supported by the Tim Gill Foundation.

References

- Manfred Bierwisch. 1989. The semantics of gradation. In *Dimensional adjectives*, pages 71–261, Berlin: Springer.
- Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *EMNLP-04*, pages 33–40, Barcelona.
- Peter Clark, William R. Murray, John Thompson, Phil Harrison, Jerry Hobbs, and Christiane Fellbaum. 2007. On the role of lexical and world knowledge in rte3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE ’07, pages 54–59, Stroudsburg, PA.
- Peter Clark, Christiane Fellbaum, Jerry Hobbs, Phil Harrison, William Murray, and John Thompson. 2008. Augmenting wordnet for deep understanding of text. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, Stroudsburg, PA.
- A. M. Collins and M. R. Quillian. 1969. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8:240 – 247.
- D. A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, New York.
- Dmitry Davidov and Ari Rappoport. 2008. Unsupervised discovery of generic relationships using pattern clusters and its evaluation by automatically generated SAT analogy questions. In *Proceedings of ACL-08, HLT*, pages 692–700, Ohio.
- James Deese. 1964. The associative structure of some common english adjectives. *Journal of Verbal Learning and Verbal Behavior*, 3(5):347–357.
- Philip Edmonds. 1999. *Semantic Representation of Near-synonyms for automatic lexical choice*. Ph.D. thesis, University of Toronto.
- Christiane Fellbaum, Derek Gross, and Katherine Miller. 1993. Adjectives in wordnet. In *Five papers on WordNet*. MIT Press, Princeton.
- Christiane Fellbaum, Peter Clark, and Jerry Hobbs. 2008. Towards improved text understanding with wordnet. In *Text Resources and Lexical Knowledge*, Berlin.
- Christiane Fellbaum. 2002. Parallel hierarchies in the verb lexicon. In *Proceedings of the Ontolex02 Workshop*, ELRA, Paris.
- Derek Gross, Ute Fischer, and George A. Miller. 1989. Antonyms and the representation of adjectival meanings. *Journal of Memory and Language*, 28:1:92 – 106.
- M. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *14th Conference on Computational Linguistics*, pages 539–545.
- Isaac Julien. 2010. Linguistic analysis with adj-scales as a tool for predicting spam product reviews. Technical report, Department of Computer Science, Princeton University.
- John S. Justeson and Slava M. Katz. 1991. Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics*, 17:1–19.
- Chris Kennedy. 2001. Polar opposition and the ontology of degrees. *Linguistics and Philosophy*, 24:33–70.
- Adam Kilgarriff. 2007. Googleology is bad science. *Computational Linguistics*, 33(1):147–151.
- Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *CoLing-98*.
- Isa Maks and Piek Vossen. 2010. Modeling attitude, polarity and subjectivity in wordnet. In *Proceedings of the 5th Global WordNet Conference*.
- Helen Moss and Lianne Older. 1996. *Word Association Norms*. Psychology Press, Hove, U. K.
- Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI-99*.
- Peter F. Schulam and Christiane Fellbaum. 2010. Automatically determining the semantic gradation of german adjectives. In *Semantic Approaches to Natural Language*, page 163, Saarbruecken.
- P.F. Schulam. 2011. Scle: A system for automatically comparing gradable adjectives. Senior Thesis.
- Vera Sheinman and Takenobu Tokunaga. 2009. Adj-scales: Visualizing differences between adjectives for language learners. *IEICE Transactions on Information and Systems*, E92-D(8):1542–1550.
- R. Snow, D. Jurafsky, and A.Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. *Advances in neural information processing systems*, 17:1297–1304.
- Peter D. Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *CoLing-08*, Manchester, UK.
- Julie Weeds and David Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475.

The C-Cat Wordnet Package: An Open Source Package for modifying and applying Wordnet

Keith Stevens and Terry Huang

Department of Computer Science
University of California, Los Angeles, 90095
`{kstevens,thuang513}@cs.ucla.edu`

David Butler

Center for Applied Scientific Computing
LLNL, Livermore 94550
`buttlerv1@llnl.gov`

Abstract

We present the C-Cat Wordnet package, an open source library for using and modifying Wordnet. The package includes four key features: an API for modifying Synsets; implementations of standard similarity metrics, implementations of well-known Word Sense Disambiguation algorithms, and an implementation of the Castanet algorithm. The library is easily extendible and usable in many runtime environments. We demonstrate it's use on two standard Word Sense Disambiguation tasks and apply the Castanet algorithm to a corpus.

1 Introduction

Wordnet (Fellbaum, 1998) is a hierarchical lexical database that provides a fine grained semantic interpretation of a word. Wordnet forms a diverse semantic network by first collecting similar words into synonym sets (*Synset*), for example “drink” and “imbibe” are connected under the verb *Synset* defined as “take in liquids.” Then, *Synsets* are connected by relational links, with the IS-A link being the most well known.

Applications typically access Wordnet through one or more libraries. Every popular programming language has at least one library: the original for C++, JWNL¹ for Java, and WordNet::QueryData² for Perl are just a few examples. While these libraries are robust and provide many features, they cannot be easily applied to two new use cases: direct modification and serialization of the database and use in a parallel processing framework, such as the Hadoop³ framework. The first has become a popular research topic in recent years, with Snow

et al. (2006) providing a well known method for adding new lexical mappings to Wordnet, and the second will increasingly become important as Wordnet applications are applied to massive web-scale datasets.

We developed the C-Cat Wordnet package to address these use cases as part of a larger information extraction and retrieval system that requires word sense information for new, domain specific terms and novel composite senses on web-scale corpora. One example includes adding new lexical mappings harvested from New York Times articles. Without support for saving additions to Wordnet and parallel processing, we would be unable to leverage existing valuable sense information. Our package solves these issues with a new API focused on modifying the database and by storing the entire Wordnet database in memory.

We designed the package to be a flexible library for any Wordnet application. It is written in Java and defines a standard Java interface for core data structures and algorithms. All code has been heavily documented with details on performance trade-offs and unit tested to ensure reliable behavior. While other Wordnet libraries exist, we hope that the release of ours facilitates the development of new, customized Wordnets and the use of Wordnet in large highly parallelized systems. The toolkit is available at <http://github.com/fozziethebeat/C-Cat>, which include a wiki detailing the structure of the package, javadocs, and a mailing list.

2 The C-Cat Wordnet Framework

Fundamentally, Wordnet acts as a mapping from word forms to possible word senses. Terms with similar senses are collapsed into a single *Synset*. The *Synset* network is then formed by linking a *Synset* to others via semantic links such as IS-A, PART-OF, and SIMILAR-TO. Our package makes two significant contributions: a collection

¹<http://sourceforge.net/projects/jwordnet/>

²<http://people.csail.mit.edu/jrennie/WordNet/>

³<http://hadoop.apache.org/>

a standardized reference implementations of well known algorithms and a new API for directly modifying and serializing the *Synset* network. Furthermore, it stores the hierarchy in memory, separating lexical queries from disc access and allows for serialization of modified hierarchies. Lastly, it provides features found in comparable libraries such as JWNL.

The C-Cat library is split up into four packages:

1. The Core API contains data format readers, writers, and *Synsets*;
2. Similarity Metrics;
3. Word Sense Disambiguation algorithms;
4. and Castanet (Stoica and Hearst, 2007), a method for automatically learning document facets using Wordnet.

2.1 Core API

The core API is centered around two interfaces: an *OntologyReader* and a *Synset*. The *OntologyReader* is responsible for parsing a Wordnet file format, building a linked *Synset* network, and returning *Synsets* based on query terms and parts of speech. The *Synset* maintains all of the information for a particular word sense, such as its definitions, examples, and links to other *Synsets*. Both interfaces provide mechanisms for modifying the sense information, *Synset* links, and lexical mappings. We store this entire structure in memory due to the minimal size of Wordnet, for example, version 3.0 is only 37 Megabytes on disk, and so that users can use Wordnet on novel distributed file systems, such as Hadoop, that do not use standard file system APIs.

Synsets are defined by three sets of values: word forms, links to other *Synsets*, and a part of speech. Each *Synset* may have multiple word forms and multiple links, but only one part of speech. We use both standard Wordnet relations and arbitrary relations to label a directed link between two *Synsets*, with the relation being stored in only the source *Synset*. We provide several methods for accessing relations and related *Synsets*: *getKnownRelationTypes()*, *allRelations()*, and *getRelations()*. In addition, each *Synset* can have a set of example sentences and a definition. To modify each *Synset*, the interface includes additive methods for relations, word forms, and examples. Furthermore, we provide a *merge()* method that takes all information from one *Synset* and adds it to another

```
OntologyReader reader = ...;
Synset cat = reader.getSynset("cat.n.1");
for (Synset rel : cat.allRelations())
    cat.merge(rel);
System.out.println(cat);
```

Figure 1: A simple merge example using the *OntologyReader* and *Synset* interfaces.

Synset. Figure 1 provides a simple example using this merge API; after the code has been run, “cat.n.1” will contain all of the information from its related *Synsets*. Lastly, the interface also permits arbitrary objects, such as ranking values, feature vectors, or additional meta data, to be attached to any *Synset* as an *Attribute*. Any *Attributes* are also merged on a call to *merge*.

OntologyReader defines an interface that maps word forms to *Synsets*. Implementations are designed to be initialized once and then used ubiquitously throughout an application. The interface provides methods for getting all *Synsets* for a word or a specific sense, for example, the query “cat.n.1” in figure 1 retrieves the first noun *Synset* for the term “cat”. To modify the sense network, we provide two key methods: *addSynset(new)* and *removeSynset(old)*. *addSynset(new)* adds a mapping from each of *new*’s word forms to *new*. *removeSynset(old)* removes all mappings from *old*’s word forms to *old*, thus removing it from the lexical mapping completely.

2.2 Similarity Metrics

While the semantic network of Wordnet is interesting on its own, many applications require sense similarity measures. As such, we provide the *SynsetSimilarity* interface that returns a similarity score between two *Synsets*. This is, in short, a Java based implementation of the Wordnet::Similarity package (Pedersen et al., 2004), which is in Perl. Figure 2 provides a naive, but short, code sample of our API that computes the similarity between all noun *Synsets* using multiple metrics.

Below, we briefly summarize the measures from (Pedersen et al., 2004) that we implemented. Several measures utilize the Lowest Common Subsumer (LCS), i.e. the deepest parent common to two *Synsets* using IS-A relations. Each measure takes in two *Synsets*, *A* and *B*, as arguments and returns a double value, typically between 0 and 1.

```

OntologyReader reader = WordNetCorpusReader.initialize(...);
Set<Synset> nouns = reader.allSynsets(PartsOfSpeech.NOUN);
SynsetSimilarity sims[] = {new PathSimilarity(), new LeskSimilarity(), ...};
for (Synset s1 : nouns)
    for (Synset s2 : nouns)
        for (SynsetSimilarity sim : sims)
            System.out.printf("%s %s %f\n", s1, s2, sim.similarity(s1, s2));

```

Figure 2: Code for computing the pairwise similarity over every noun *Synset* using multiple metrics

Path Based Methods measure the similarity based on a path connecting *A* and *B*. *Path* simply returns the inverse length of the shortest path between *A* and *B*. *Leacock&Chodorow* (Leacock and Chodorow, 1998) returns the length of the shortest path scaled by the deepest depth in the hierarchy. *Wu&Palmer* (Wu and Palmer, 1994) returns the depth of the LCS scaled by the cumulative depth of *A* and *B*. *Hirst&St.Onge* (Hirst and St-Onge, 1997) uses all links in the hierarchy and measures the length of the path that is both short and has very few link types.

Lexical methods measure the amount of lexical overlap between *A* and *B*. *Lesk* (Lesk, 1986) returns the number of words overlapping in *A* and *B*'s glosses. *ExtendedLesk* (Banerjee and Pedersen, 2003) extends Lesk by also comparing the glosses between any *Synsets* related to *A* or *B*.

Information based Methods utilize the Information Content (IC) of a *Synset*, which measures the specificity of the terms in a *Synset* as measured in a sense tagged corpus. *Resnick* (Resnick, 1995) returns the IC of the LCS. *Jiang&Conrath* (Jiang and Conrath, 1997) returns the inverse difference between the total IC of *A* and *B* and the IC of their LCS. *Lin* (Lin, 1998) returns the IC of the LCS scaled by the total IC of *A* and *B*.

In addition to the raw similarity metrics, we provide a utility classes that return meta information about a pair of *Synsets* such as their shortest path, their LCS, and several other helpful methods.

2.3 Word Sense Disambiguation

Word Sense Disambiguation is perhaps the most standard application of Wordnet. Disambiguation models attempt to select a *Synset* for a given word that best matches a given context. For example, an algorithm might select the river bank *Synset* of “bank” for the context “he sat on the bank of the river” rather than the financial institution *Synset*. We provide a

WordSenseDisambiguation interface that applies word sense labels to tokenized sentences. Currently, we only provide a small number of unsupervised algorithms, but plan on adding more. Below, we briefly describe each algorithm.

Lexical Methods rely on lexical information in Wordnet to disambiguate words. *Lesk* (Lesk, 1986) selects the *Synset* that has the highest total Lesk similarity to the *Synsets* for other context words. *ExtendedLesk* (Banerjee and Pedersen, 2003) extends Lesk by using the Extended Lesk similarity metric for all comparisons. *MostFrequentSense* selects the first *Synset* returned by Wordnet for a given term. This serves as a canonical baseline which is often challenging to outperform.

Graphical Methods treat the network as a graph and disambiguate using a number of measurements. *PersonalizedPageRank* (Agirre and Soroa, 2009) (*PPR*) runs the PageRank algorithm over an undirected graph composed from the entire Wordnet network. Words needing disambiguation are given “artificial” nodes that link to their possible *Synsets*. For each ambiguous word, the algorithm selects the highest ranking *Synset*. *DegreeCentrality* (Navigli and Lapata, 2010) (*DC*) forms a subgraph from the Wordnet network composed of ambiguous content words in a sentence and the *Synsets* that connect their possible *Synsets*. It assigns to each word the target *Synset* with the highest degree in the subgraph. *PageRankCentrality* (Navigli and Lapata, 2010) (*PRC*) composes the same subgraph as *DegreeCentrality*, but performs PageRank on this subgraph and selects the *Synset* with the highest rank for each ambiguous word.

2.4 Castanet

Amazon.com and other online retailers often display manually crafted facets, or categories, for product navigation. A customer can start browsing from the Book category and dive down into more

specific categories such as Fiction, Entertainment, or Politics. These facets form a hierarchy of categories and each category is subdivided until a narrow set of interesting items are found. Unfortunately, not all datasets have well structured meta data. The Castanet algorithm automatically learns this hierarchical faceted meta data (HFC) for a set of documents by using discriminative keywords (Stoica and Hearst, 2007), making structured navigation possible for arbitrary document sets.

Castanet takes advantage of Wordnets IS-A hierarchy to automatically create HFC. Castanet first extracts keywords from the set of documents (we use term-frequency inverse document frequency, TF-IDF, by default, but our API allows for other methods). For each extracted keyword, Castanet then creates a chain of words that lead from the root of the hierarchy to the keyword’s *Synsets*. Each keyword chain is then merged together to form a “backbone” tree which is later reduced by eliminating redundant or non-discriminative nodes, such as those with one child.

Our Castanet API is both simple and flexible. To create a Castanet tree, one calls *Castanet.buildTree()* with a directory path to a set of text documents. Our implementation will automatically extract keywords, extract the backbone tree, and finally index each document under its learned facets. The returned result allows users to fully navigate the documents via the learned facets. We also provide an example Java web service for exploring the hierarchy in a browser.

3 Benchmark

To evaluate our library, we apply our six WSD implementations against two standard evaluations: the all words disambiguation tasks from SenseEval 3 (Snyder and Palmer, 2004) and SemEval 2007 (Pradhan et al., 2007), these use Wordnet version 1.7.1 and 2.1 respectively. We answer all test instances except those that do not have any mapping in Wordnet. Before processing, we apply part of speech tags to each token using the Open NLP MaxEnt Tagger ver 1.5.0⁴. We use the original databases as a baseline, called Base, in our experiments and test our modification API by adding the eXtended Wordnet (XWN) relations (Mihalcea and Moldovan, 2001) to each database and disam-

biguate using these extended Wordnets⁵.

Model	Ver	SenseEval-3	SemEval-07
MFS	Base	59.8	49.4
Lesk	Base	35.2	27.7
E-Lesk	Base	47.8	37.6
PPR	Base	42.9	32.8
DC	Base	43.2	33.3
PRC	Base	31.7	22.7
Lesk	XWN	35.2	27.7
E-Lesk	XWN	39.9	33.9
PPR	XWN	50.3	36.7
DC	XWN	47.3	37.1
PRC	XWN	33.0	24.0

Table 1: F1 Word Sense Disambiguation scores on the two test sets

Table 1 presents the F1 score for each algorithm using the original and extended databases. As expected, the MFS baseline outperforms each unsupervised algorithm. Although our scores do not match exactly with previous publications of these algorithms, we still see similar trends and the expected gains from adding new relations to the hierarchy. For *DegreeCentrality* and *PageRankCentrality*, our different results are likely due to a implementation difference: when extracting a subgraph from Wordnet, we only use directed links as opposed to undirected links for computational efficiency. Other variations are possibly due to different methods of handling multi-word expressions and our part of speech tags. Still, *DC* gains about 4% points with XWN relations and *PPR* gains about 7% points on Senseval-3. Unexpectedly, *ExtendedLesk* actually does worse with the additional relations.

We also performed a visual test of our Castanet implementation. We ran the algorithm over 1,021 articles extracted from the BBC World News using Wordnet 3.0. The articles came from a diverse set of categories including world, business, technology, and environmental news. Figures 3 and 4 show snapshots of our Castanet web application. Figure 3 displays the top level facets displayed to a new user. The top bar of this screen can break down the facets alphabetically to facilitate facet selection. Figure 4 shows a snapshot of several documents found after selecting several facets. It displays the selected facets, document titles, document text, and interesting key words. While this is only a simple interface, it provides an example of what our implementation can accomplish and how

⁴<http://opennlp.sourceforge.net/models-1.5/>

⁵Note that we added XWN 1.7 relations to Wordnet 1.7.1 and XWN 2.0 relations to Wordnet 2.1, some links were discarded due to updates in Wordnet.

to use our API.

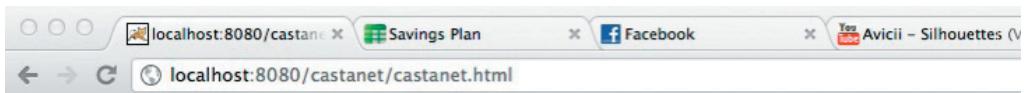
4 Future Work and Conclusion

We have presented our Java Wordnet library that provides two new key features: maintenance of an in memory database and an API centered around modifying the network directly. Additionally, we've provided implementations of several well known similarity metrics, disambiguation algorithms, and the Castanet information retrieval algorithm. All code is unit tested, heavily documented, and released under the GPL Version 2 license. We are currently working to extend our newest APIs, such as those for WSD and Castanet, to handle more interesting use cases. In the future work we hope to expand this library with an evaluation framework for customized Wordnets, such as those generated by (Snow et al., 2006).

Acknowledgements This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-CONF-499791).

References

- Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Satyanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th international joint conference on Artificial intelligence*, pages 805–810, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, illustrated edition edition, May.
- G. Hirst and D. St-Onge. 1997. Lexical chains as representation of context for the detection and correction malapropisms.
- J.J. Jiang and D.W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, pages 19–33.
- C. Leacock and M. Chodorow. 1998. *Combining local context and WordNet similarity for word sense identification*, pages 305–332. In C. Fellbaum (Ed.), MIT Press.
- Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.
- Dekang Lin. 1998. An information-theoretic definition of similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann.
- Rada Mihalcea and Dan I. Moldovan. 2001. extended wordnet: progress report. In *In Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pages 95–100.
- Roberto Navigli and Mirella Lapata. 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:678–692, April.
- Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL—Demonstrations '04, pages 38–41, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Sameer S. Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, srl and all words. In *In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007*, pages 87–92.
- Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453.
- Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, ACL-44, pages 801–808, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July. Association for Computational Linguistics.
- Emilia Stoica and Marti A. Hearst. 2007. Automating creation of hierarchical faceted metadata structures. In *In Procs. of the Human Language Technology Conference (NAACL HLT)*.
- Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.



Current Path: abstraction

[ALL](#) [0-9](#) [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

- [previous](#)
- [report](#)
- [psychological_feature](#)
- [part](#)
- [measure](#)
- [group](#)
- [attribute](#)

Figure 3: A sample view of learned Castanet facets for the BBC Word News data set

Current Path: abstraction > part > compound > oil

[ALL](#) [0-9](#) [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

- [previous](#)

/Users/thuang513/research/C-Cat/extendOntology/test-docs/LibyaNews.txt
keywords: oil,water

With his eccentric , often inscrutable personality , Kadaffi has ruled Libya , one of the world 's largest oil producers , for 41 years through a mix of repression , patronage and shrewd tribal alliances .

Two of the Arab world 's most ruthless leaders have moved to crush revolts threatening their power in Libya and Yemen as security forces and thugs intensified attacks on dissidents and protesters dug scores of fresh graves amid the rattle of gunfire .

Saleh , who once described ruling Yemen as " dancing on the heads of snakes " , has stayed in power for 32 years in much the same manner .

/Users/thuang513/research/C-Cat/extendOntology/test-docs/LibyaNews2.txt
keywords: oil,water

The chaos that has consumed Libya since protesters last week began pushing for Col. Gadhafi 's ouster has spawned an array of security concerns - over oil supplies , the safety of tens of thousands of foreign workers there and the risks posed by the weapons in Col. Gadhafi 's remaining arsenal .

Prices for light , sweet crude for April delivery - the main U.S. oil contract - at one point in the trading day hit \$ 100 a barrel for the first time in more than two years .

Oil prices surged over fears about the security of supplies from Libya , a major oil producer .

/Users/thuang513/research/C-Cat/extendOntology/test-docs/TaxGas.txt
keywords: oil,water

In a news conference , Obama said that a payroll tax cut signed into law in December as part of the tax package would now go to cushion the impact of a recent spike in oil prices and allow for continued economic growth .

Obama pushed back against both arguments Friday , saying the pickup in oil prices is largely being driven by the global economic recovery , particularly the speed of growth in emerging economies , and does n't yet necessitate tapping the emergency reserve .

President Obama acknowledged Friday that the fast-rising cost of gasoline could diminish the effect of policies designed to stimulate economic growth , but warned that he is not yet prepared to unleash the nation 's energy reserves to bring down the price of oil .

Figure 4: A sample view of a discovered document after navigating the Castanet facets

Linking WordNet to DBpedia

Aynaz Taheri

Computer Engineering Dept., Shahid
Beheshti University, Tehran, Iran
ay.taheri@mail.sbu.ac.ir

Mehrnoosh Shamsfard

Computer Engineering Dept., Shahid
Beheshti University, Tehran, Iran
m-shams@sbu.ac.ir

Abstract

In this paper we present the process of matching two important datasets in Linking Open Data (LOD): DBpedia and WordNet 3.0. DBpedia plays the main role in the LOD cloud. It is an influential knowledge base and consists of over one billion pieces of information about millions of things. On the other hand Princeton WordNet is the most important lexical ontology. WordNet 3.0 in RDF is also one of the resources in Linking Open Data. Certainly, linking these two common datasets has impressive effects in various aspects of consuming them. In this paper the methodology of matching, statistical information and some beneficial use cases of the matching is explained.

1 Introduction

Nowadays increasing the amount of linked data in Linking Open Data project is not the only challenge of publishing linked data; rather, mapping and linking the linked data resources are also equally important and can improve the effective consuming of linked data resources. Without these links, we confront with isolated islands of datasets, which could not exploit knowledge of each other. The fourth rule of publishing linked data in (Bizer, Heath, et al., 2009) explains the necessity of linking URIs to each other. Therefore, extension of datasets without interlinking them is against the Linked Data principles. The importance of this issue increased our motivation of doing mapping between two core datasets of Linking Open Data.

DBpedia is a significant knowledge base.

DBpedia knowledge extraction framework extracted its knowledge from Wikipedia and converted itself as a crystallization point for the web of data (Bizer, Lehmann, et al., 2009). DBpedia currently has knowledge for more than 3.6 million things about persons, places, music, films, video games and etc (“DBpedia,” n.d.). It also contains information for these things in different languages. Most of the dataset publishers try to link their datasets to DBpedia. In (Cyganiak, 2010) you can see the mass of links to DBpedia. Some of links from DBpedia to these datasets are available in (“Interlinking DBpedia,” 2011) and one of these datasets is WordNet (W3C).

WordNet (Fellbaum, 1998) is an electronic lexical database that is designed in Princeton University for English language. WordNet uses synonymous sets, called synset. The latest version of WordNet contains 155,287 words organized in 117,659 synsets (“WordNet 3.0 database,” n.d.). WordNet includes nouns, adjectives, verbs and adverbs. Synsets in WordNet are connected to each other with semantic relation such as: synonymy, antonymy, hyponymy, hypernymy, meronymy, troponymy and etc.

Lexical ontologies like WordNet are important resources in natural language processing (NLP). They are used in various tasks and applications, especially where semantic processing is evolved such as question answering, machine translation, text understanding, information retrieval and extraction, knowledge acquisition and semantic search engines (Shamsfard, 2008). Integration of Princeton WordNet and DBpedia could improve the semantic processing. Princeton Wordnet has been mapped to most of the WordNets developed for other languages in the world. So, WordNets of these languages could be linked to DBpedia via Princeton WordNet and the result of WordNet to DBpedia matching will affect NLP in different languages.

At the present time, WordNet is available in

the Linking Open Data cloud. There are two datasets in the LOD cloud which represents WordNet in the form of linked data. One of them is WordNet (W3C¹) (Assem et al., 2006) that is the OWL/RDF representation of Princeton WordNet 2.0 and the other one is WordNet (VUA²) (“Wordnet 3.0 in RDF” 2010) that is the RDF version of WordNet 3.0. WordNet (VUA) is mapped to WordNet (W3C) and DBpedia is also linked to WordNet (W3C).

Each synset in WordNet (VUA) has an URI. Synsets are derefencable by their URIs and via HTTP protocol. Instances of synsets have also URIs. There are specified patterns for URIs of synsets and instances. The word “instance” is depicted in URIs of instances.

Currently, DBpedia has 467101 links to WordNet (W3C). But there are shortcomings in these links. We are going to cover these defects in our matching:

1. There are only hypernymy relations from instances of DBpedia to noun synsets of WordNet in current links and there is no relation from WordNet to DBpedia. It is considerable that these relations only represent a kind of instantiation. An example of these relations is following:
<http://dbpedia.org/resource/White_House>
<http://dbpedia.org/property/wordnet_type>
<<http://www.w3.org/2006/03/wn/wn20/instances/synset-building-noun-1>>

In the above example, it is demonstrated that ‘White_house’ is an instance of ‘building’ synset in WordNet and nothing more. Whereas, WordNet has information about ‘white_house’. There is a synset in WordNet 2.0 with this URI: “http://www.w3c.org/2006/03/wn/wn20/instances/synset-White_house-noun-2”. Matching the WordNet synset and the correspondent one in DBpedia is desirable for us. There are many synsets in WordNet which have equivalents in DBpedia. Discovering this type of relations is one of our motivations for doing this project.

2. There is another kind of relation that detecting it between WordNet and DBpedia is beneficial. We find instantiation or hypernymy relations between noun synsets of WordNet and DBpedia classes. It is important to know a synset of WordNet belongs to which concept from the viewpoint of DBpedia.

3. There are many properties in DBpedia. There is no link from these properties to equivalents in WordNet.

4. Only noun synsets of WordNet are considered in current links of DBpedia to WordNet. We are going to find equivalents of verb and adjective synsets in WordNet to properties in DBpedia too.

The rest of the paper is organized as follows: Section 2 presents our methodology of matching WordNet 3.0 in RDF to DBpedia. Section 3 explains statistical information about the result of mapping. Section 4 describe some use cases and advantages of the mapping WordNet3.0 to DBpedia. Section 5 discusses evaluation of the results and section 6 provides some conclusion about this paper.

2 Methodology of Matching

We are going to find coreferent URIs in WordNet (VUA) and DBpedia. This process is also known as entity matching, object resolution, object consolidation, entity identification, identity recognition, identity disambiguation or instance matching. In recent years many efforts have been done for making tools, softwares and frameworks for detecting coreferent URIs. One of these important products is Silk framework (Volz et al., 2009). Silk is a link discovery framework for the web of data that uses a declarative language (Silk-LSL) for specifying which types of links should be found between which types of entities. DBpedia has counseled utilizing this tool for generating links from other datasets to DBpedia (“Interlinking DBpedia,” 2011). In our work, we do not use Silk because it needs all kinds of relations that might be discovered between entities to be described by user beforehand. We instead, apply our approach for generating links between two datasets.

Our methodology consists of two main phases: preliminary, supplementary.

1. Preliminary Phase:

In this phase, we use a terminological method for comparing synsets in WordNet and entities in DBpedia. The terminological method is applied in three steps:

- *Matching instances of WordNet to instances and classes of DBpedia:*

At the first step, equivalent instances in WordNet and DBpedia are found. Instances’ synsets in WordNet (UVA) are available apart from other noun synsets. Thus, in the first step we discover all equivalent URIs from these two sets. After finding these equivalences, the next step is to detect the correspondences between Wordnet instances and DBpedia

¹ World Wide Web Consortium

² Vrije University Amsterdam

classes.

There are instantiation relations between instances and their classes in DBpedia. In fact, the types of instances are described with these relations. This relation in DBpedia is represented with:

"<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>". Since, we found equivalences relations between instances in WordNet and DBpedia, and on the other hand there are instantiation or hypernymy relations between instances and their classes in DBpedia, so it is possible to represent the type of WordNet instances in DBpedia. Figure 1 indicates this process.

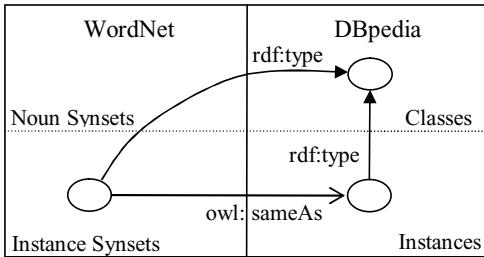


Figure 1. Process of matching at the first step in the first phase

- *Matching noun synsets of WordNet to classes in DBpedia:*

Some noun synsets of WordNet have equivalents in classes of DBpedia. For example, both of the datasets have knowledge about "Language". So, the "synset-Language-noun-1" synset in WordNet is the same as "Language" class in DBpedia.

<http://purl.org/vocabularies/princeton/wn30/synset-language-noun-1>
<http://www.w3.org/2002/07/owl#owl:equivalentClass>
<http://dbpedia.org/ontology/Language>

- *Matching noun, verb and adjective synsets of WordNet to properties in DBpedia:* The aim of this phase is to recognize properties of DBpedia and their equivalents in WordNet. Properties play the predicate role in a triple. So, detecting equivalent synsets with properties is advantageous. There are three kinds of properties in DBpedia: owl:ObjectProperty, owl:DatatypeProperty and property. owl:ObjectProperty and owl:DatatypeProperty are properties in ontology of DBpedia but the third kind of properties are independent from DBpedia ontology and there are no structural and hierarchical relations between them ("The DBpedia Data Set," 2011). These properties

are created directly from Wikipedia infobox properties with no regard to DBpedia ontology. There are some properties in these three kinds that seem to be equivalent. The next example represents that there are two properties in DBpedia that specify "Language" property: <http://dbpedia.org/ontology/language> : an object property

<http://dbpedia.org/property/language> : a property

In all of the three steps, similarity computing method is a token-based distance computing. In this method a string is considered as a bag of words (Euzenat,2007). In our matching method the label of entities in DBpedia, the label of the WordNet synsets in their URIs, the label of senses in a synset and the description (gloss) of a synset in WordNet are transformed to bags of words. But before this transformation we normalize strings and remove stop words. After producing the bags of words, a measure for estimation of similarity between WordNet synset and DBpedia entity is applied (1).

X: a bag of words including words in the label of synset

Y: a bag of words including words in the label of DBpedia entity or its comment

S: a bag of words including words in the label of senses of synset

G: a bag of words including words in the glossary of synset

$$\delta(X, Y) = \frac{|X \cap Y| + |Y \cap (X \cup S \cup G)|}{|X| + |Y|} \quad (1)$$

2. Supplementary Phase:

Similar entities regarding lexical features were found in the previous phase. So, we are sure that the matching results of the first phase are lexically correct. But these results are not accurate necessarily and maybe there are correspondences that don't have entities with the same identity and there are only lexical similarities between them. The purpose of this phase is to refine the result of matching in the previous phase. In other words, a method for URI disambiguating is described in this phase. We used hierachal structure of WordNet and DBpedia for disambiguation.

'wnschema:instanceOf' and 'wn20schema:hyponymOf' relations are used for gaining an understanding of taxonomic structure in WordNet. 'wnschema:instanceOf' relation denotes a relation between an instance synset and a noun synset. Two relations from DBpedia are also utilized for determining the taxonomic structure. These two relations are 'rdf:type' and

‘rdfs:subClassOf’. The first one denotes a relation between an instance and the class that belong to and the second one expresses a relation between two classes. These relations are used as sources for disambiguation.

Some structure based techniques are presented in (Euzenat and Shvaiko, 2007). One of them is Wu-Palmer similarity measure (Wu and Palmer, 1994). This similarity measure is used in the second phase of our methodology. Consider the following URIs:

- URI1: <http://purl.org/vocabularies/princeton/wn3/0/synset-Pluto-noun-1>
- URI2: <http://purl.org/vocabularies/princeton/wn3/0/synset-Pluto-noun-2>
- URI3: <http://purl.org/vocabularies/princeton/wn3/0/synset-Pluto-noun-3>
- URI4: <http://dbpedia.org/resource/Pluto>

URI1, URI2 and URI3 from WordNet are lexically similar to each other and the first phase finds them equal to URI4 from DBpedia. While, URI1 has ‘wnschema:instanceOf’ relation with ‘fictional_character’ and is a cartoon character. URI2 has ‘wnschema:instanceOf’ relation with ‘Greek_deity’ and is the god of the underworld in ancient mythology. URI3 has ‘wnschema:instanceOf’ relation with ‘outer_planet’ and is a small planet. URI4 has ‘rdf:type’ relation with ‘Planet’ class of DBpedia ontology. The matching of URI1 and URI2 with URI4 is obviously wrong.

We use Wu-Palmer measure and assess the similarity of taxonomic structure of URIs in WordNet. In the former example, the similarity of (fictional_character, Planet) is 0.48 and the similarity of (Greek_deity, Planet) is 0.46. These similarities are less than our threshold. Therefore, the matching of URI1 and URI2 with URI4 are excluded from the results.

All of the correspondences with structural similarity less than the threshold are removed from the result of matching.

The result of matching is available at:
<http://step1.nlplab.sbu.ac.ir/wordnetdbpedia/matching.aspx>

Subject(WordNet)	Predicate	Object(DBpedia)	Number of Matching
Instances	owl:SameAs	Instances	27923
Instances	rdf:type	Classes	18555
Noun, Verb, Adjective	owl:equivalentProperty	Object Property	583
Noun, Verb, Adjective	owl:equivalentProperty	DataType Property	438
Noun, Verb, Adjective	owl:equivalentProperty	Property	10379
Noun	owl:equivalentClass	Classes	344

Table 1. Result of Matching

3 Statistical Information of the Matching Outcome

In table 1 the result of WordNet to DBpedia is represented. In the first column, the kinds of synsets are denoted. In the second column the kinds of relations that are discovered and in the third column elements of DBpedia in matching are represented.

4 Use Cases

Influences of the matching consequences are clearly perceptible in the Natural Language Processing and Semantic Processing domains. We categorize use cases of WordNet 3.0 to DBpedia matching in three groups:

- *Enriching WordNet*: Princeton WordNet can be enriched with more relations. Properties in DBpedia can be used for finding more relations in Wordnet.
- *Developing Formal Ontologies*: Princeton WordNet is a lexical ontology and is far from a formal ontology. The types of relations in WordNet are restricted to synonymy, antonymy, hyponymy, hypernymy, meronymy, troponymy. For moving toward a formal ontology, it is necessary to augment the relations between synsets. With utilizing interlinking of WordNet and DBpedia, it is possible to discovering relations between synsets of WordNet via their correspondent entities in DBpedia. Due to the fact that DBpedia is an important knowledge base and have information for entities in the form of properties. These properties can be exploited for making WordNet a formal ontology.
- *Semantic Search*: In semantic search, disambiguating words is a main challenge. Taking advantage of WordNet to DBpedia matching, could help disambiguating through the large amounts of information about entities and properties in DBpedia.
- *Finding more instances for WordNet synsets*: DBpedia is greater than WordNet in the number of instances. We discovered equivalent relations

between some synsets of WordNet and classes of DBpedia. So we can apply the instances of equivalent class for the instantiation of the synset.

5 Evaluation

We evaluated a subset of matching result manually. This subset contained 500 members. The value of obtained precision is 0.92. Computation of recall is not possible for this project as there is no golden standard and manual extraction of all possible links between these two sets is almost impossible.

6 Conclusions and Future Work

In this paper we discussed about matching two important datasets in LOD cloud: DBpedia and WordNet. Shortcomings about the current links from DBpedia to WordNet presented and the necessity of generating more different kinds of links between them is explained. Interlinking these two datasets can improve applications on natural language processing and semantic processing; furthermore WordNet is also impressionable from the matching result and can move toward an enriched lexical ontology or even a formal ontology.

Future work will focus on mapping WordNets of other languages especially those with less resources to linked data. Linking WordNets to DBpedia is possible via the outcome of Princeton WordNet to DBpedia matching. For example FarsNet, the Persian wordnet (Shamsfard, et al., 2010) is a good candidate for this mapping. FarsNet to Princeton WordNet mapping is available, so matching FarsNet to DBpedia is possible. After linking FarsNet to DBpedia, we are going to extend relations in FarsNet with utilizing DBpedia. Accordingly, FarsNet will be connected to LOD cloud and causes improvements in Persian language semantic processing.

References

- Mark V. Assem, Aldo Gangemi and Guus Schreiber. 2006. Conversion of WordNet to a standard RDF/OWL representation. In proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy.
- Mark V. Assem, Aldo Gangemi and Guus Schreiber. RDF/OWL Representation of WordNet. (2006). Retrieved April, 2011, from <http://www.w3.org/TR/wordnet-rdf/>
- Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soren Auer, Christian Becker, Richard Cyganiak and Sebastian Hellmann. 2009. DBpedia- A crystallization point for the Web of Data. *J. Web Sem.* 7(3): 154-165.
- Christian Bizer, Tom Heath and Tim Berners-Lee. 2009. Linked Data-The Story So Far, *Int. J. Semantic Web Inf. Syst.* 5(3) :1-22.
- Richard Cyganiak and Anja Jentzsch. The Linking Open Data cloud diagram. (2010). Retrieved April, 2011, from <http://richard.cyganiak.de/2007/10/lod/>
- DBpedia. (2011). Retrieved April, 2011, from [http://dbpedia.org/About](http://dbpedia.org/) .
- Jerome Euzenat and Pavel Shvaiko. 2007. *Ontology matching*. Springer-Verlag, Berlin Heidelberg.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Interlinking DBpedia with other Data Sets. (2011). Retrieved April, 2011, from <http://wiki.dbpedia.org/Interlinking>.
- Mehrnoosh Shamsfard, Akbar Hesabi, Hakimeh Fadaie, Niloofar Mansoory, Ali Famian, Somayeh Bagherbeigi, Elham Fekri, Maliheh Monshizadeh, S. Mostafa Assi. 2010. Semi Automatic Development of FarsNet: The Persian WordNet, In Proceedings of the 5th Global WordNet Conference, Mumbai, India.
- The DBpedia Data Set. (2011). Retrieved April, 2011, from <http://wiki.dbpedia.org/Datasets>.
- Julius Volz, Christian Bizer, Martin Gaedke, Georgi Kobilarov. 2009. Silk – A Link Discovery Framework for the Web of Data. In Proceedings of the 2nd Workshop about Linked Data on the Web. Madrid, Spain.
- WordNet 3.0 database statistics. Retrieved April, 2011, from <http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html>.
- Wordnet 3.0 in RDF. (2010). Retrieved April, 2011, from <http://semanticweb.cs.vu.nl/ld/wn30/>
- Zhibiao Wu and Martha Stone Palmer. 1994. Verb Semantics and Lexical Selection. In proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL), 133–138, Las Cruces.

Extension of Phrases for Article Determination using WordNet Thesaurus

Hiromi Takeuchi¹, Hirofumi Miyake¹, Atsuo Kawai¹, Ryo Nagata², Hokuto Otake³

¹Mie University, Tsu, Japan

²Konan University, Kobe, Japan

³Fukuoka University, Fukuoka, Japan

{ takeuti, miyake, kawai } @ai.info.mie-u.ac.jp

nagata-gwa2012@hyogo-u.ac.jp

ototake@fukuoka-u.ac.jp

Abstract

We propose a statistical model for detecting and correcting article errors, which are often made by Japanese when writing in English. To solve the data sparseness problem in conventional phrase-based statistical models, we proposed two approaches. In the first approach, we used a statistical test to efficiently select and apply reliable phrases for article determination. In the second approach, we proposed to extend phrases for article determination by using the hypernyms of the WordNet thesaurus. In addition, we selected relevant hypernyms when multisense words, which have multiple meanings, have multiple hypernyms. Experiments showed that our proposed method outperformed the conventional method. The F-measure at $\theta = 0.95$, which has the highest precision but the lowest recall point, shows 12.4% improvement compared to the method without using the thesaurus and the statistical test.

1 Introduction

Opportunities for Japanese to write in English are increasing as importance of English as an international communication language continues to increase. Writing grammatically correct English is very difficult, however, because the Japanese language does not use articles: *a*, *an*, *the*, ϕ (hereinafter ϕ denotes no article). Therefore Japanese often make article errors (Kawai et al., 1984; Izumi et al., 2003). On the other hand, the English article system is crucial because it provides such important information as the restriction of noun phrases.

Article errors often cause misunderstanding of intended meaning especially in documents that require a clear context such as technical papers. To correct article errors, many Japanese have their papers proofread by native speakers of English who have specialized knowledge. The need for grammatical error detection and correction systems is increasing because human proofreading and rewriting are time-consuming and expensive.

Researchers have studied methods for detecting and correcting article errors. Most of the work uses machine learning-based classifiers to detect article errors and/or to determine. Knight and Chander (1994) takes the first step to using a machine learning algorithm for determining the choice between *a* and *the*. Minnen et al. (Minnen et al., 2000) extend this work to three-way classification (*a*, *the*, and ϕ). Izumi et al. (2003) and Han et al. (2004) propose maximum entropy (ME) classifier-based methods for predicting correct articles. The features are based on lexical and syntactic information around the article in question. Han et al. report that their method achieves a recall of 0.40 with a precision of 0.90.

Although the classifier-based methods achieve good performances, they have a drawback. It is difficult to know why the classifier chooses the article to use. In other words, one cannot easily interpret which features contribute to the determination to what extent. In terms of writing support, it is crucial to instruct the user why the article is appropriate in the context in question. Only then, the user can examine whether he or she rewrite his or her writing or not.

To overcome the drawback of the classification-based methods, Nagata et al. (Aug.2005) proposed a phrase-based method for article determination. The method tries to find meaningful phrases such as idioms for determining articles based on statistics obtained from a given corpus. The extracted phrases are used to determine which article to use as well as to give the feedback to the user. For example, the phrase-based method may extract a phrase such as *an accuracy of* to correct *The method achieved φ accuracy of 0.700.* from a corpus consisting of technical papers.

An inherent drawback to the phrase-based method is that they severely suffer from the data sparseness problem. Namely, it is often the case that there is not sufficient data (and statistics) to obtain reliable phrases no matter how large the corpus is, which leads to undesirable false negatives and false positives.

In this paper, we propose two approaches to overcome the data sparseness problem. In the first approach, we use a statistical test to efficiently select and apply reliable phrases for article determination. It is done without human intervention and is purely calculated from statistics extracted from a corpus. In the second approach, we propose to expand phrases for article determination by using semantic classes. Intuitively, the data sparseness problem is expected to be reduced by replacing words with their corresponding semantic class in obtaining phrases for article determination. For example, instances of *eat φ breakfast* and *eat φ lunch* in a corpus would both be replaced by *eat φ [meal#n#1]*, which results in obtaining a new phrase and increasing statistics.

The use of semantic class brings up an additional advantage. In textbooks on English grammar, article usage is often explained based on semantic classes (e.g., “If an unmodified means-of-transport noun, such as *bus*, follows preposition *by*, then it will have a null determiner, that is, ϕ (Bond, 2005)”). Considering this, the expansion based on semantic classes make it easier that the user interprets the feedback. Also, the user may come up with a new phrase which does not appear in the training corpus from the feedback. For example, one may come up with *eat φ dinner* in the association with *eat φ [meal#n#1]*.

Unfortunately, however, simply replacing words with their corresponding semantic class introduces a

new problem. That is, a word may belong to more than one semantic class. This means that one has to determine by which semantic class the word in question should be replaced when expanding phrases for article determination. This paper also proposed an effective method for solving this problem.

The plan of this paper is as follows. Section 2 describes the phrase-based method. Section 3 describes the proposed method to overcome the data sparseness problem. Section 4 describes experiments about conventional phrase-based method and proposed method. Finally, Section 5 describes conclusions and discussions.

2 Phrase-based Method

2.1 Phrase for article determination

The phrase-based method has five steps. In the first step, the phrases are extracted from the obtained Key Word In Context (KWIC). Here, because the purpose is to make phrases for article determination, keyword becomes one of the article (*a(n)*, *the*, ϕ). A phrase is defined as a sequence of words that consists of the keyword, n_+ words following the keyword, and n_- words preceding it. In other words, a phrase can be regarded as an n-gram centered on the keyword. For example, the following KWIC:

... level vi covers ϕ wheat entered after ...

gives a set of phrases:

vi	covers	ϕ	wheat	entered
vi	covers	ϕ	wheat	
vi	covers	ϕ		
	covers	ϕ	wheat	entered
	covers	ϕ	wheat	
	covers	ϕ		

when $n_+ \leq 2$ and $n_- \leq 2$. In this paper, the values of n_+ and n_- are set to $n_+ \leq 5$ and $n_- \leq 5$, respectively, on the assumption that the length of most collocations does not exceed 11 words.

In the second step, all phrases extracted from the KWIC are sorted alphabetically. If a certain phrase appears twice or more in the KWIC, identical phrases are repeated in sequence.

In the third step, the same phrases are merged into one and their frequencies are counted. For example,

the phrases in the above example give (phrase; its frequency):

covers	ϕ	wheat	; 15
	ϕ	wheat	; 249
	:	.	

From the merged phrases, determining each phrase's frequency is easy.

In the fourth step, the conditional probabilities of the merged phrases for article determination is calculated.

Now, we introduce some variables and symbols to formalize the extraction procedure. Variable A denotes the keywords (one of the articles) in the KWIC. Variable C denotes an n-gram whose keyword is any value of the articles (e.g., $C = \{\text{a(n)/the}/\phi\}$ wheat"). Symbols f and p denote frequency and probability. The conditional probability can be estimated by

$$p(A|C) = \frac{f(A,C)}{f(C)}. \quad (1)$$

In the fifth and final step, a phrase for article determination is used that has the highest probability in the applicable phrase. If the highest probability is less than θ , we do not generate articles. Here, θ ($0 \leq \theta \leq 1$) is set by the system users depending on the intended use.

3 Proposed Method

3.1 Proposed Method using Reliability of Phrases for Article Determination

In Eq. (1), if the phrase's frequency for article determination is small, the degree of the bias of the distribution of articles is not reliable. For example, assume one phrase which appears only once and co-occurs only with article α once (probability = 1.0). At the other extreme, assume another phrase which appears 100 times and co-occurs with an article α 95 times (probability = 0.95). Intuitively, the latter phrase for article determination seems more reliably biased. To evaluate the statistical significance of the bias, we propose the χ^2 test, which is very common for evaluating a bias between expected and observed frequencies. For each phrase, the frequency of co-occurrence with an article is regarded as a sample

value; the null hypothesis is that the occurrence of each article frequency $f(C)$ is independent from the occurrence of phrases for article determination. We expect to reject this hypothesis.

We denote $EP(A)$; the expected probability (under the null hypothesis) of each article. Here, $EP(A)$ always become 0.33. Because, under the null hypothesis, articles (a(n), the, ϕ) randomly occur. The statistical value of χ^2 is defined as follows:

$$\chi^2(C) = \begin{cases} \sum_{A \in \{\text{a(n), the, } \phi\}} \frac{(f(A,C) - f(C) \times EP(A))^2}{f(C) \times EP(A)} & (f(C) > 15) \\ \sum_{A \in \{\text{a(n), the, } \phi\}} \frac{(|f(A,C) - f(C) \times EP(A)| - 0.5)^2}{f(C) \times EP(A)} & (f(C) \leq 15) \end{cases} \quad (2)$$

For example, if $f(\text{"covers a wheat"}) = 2$, $f(\text{"covers the wheat"}) = 3$ and $f(\text{"covers } \phi \text{ wheat"}) = 95$, then $f(C) = f(\text{"covers }\{\text{a(n), the, } \phi\}\text{ wheat"})$ becomes 100. And $\chi^2(C)$ becomes 172.88 by Eq. (2).

If $\chi^2(C) > \chi_\beta^2$, the null hypothesis is rejected with significance level β , and means n-gram C is reliably biased from the view of the article distribution. And the most frequent phrase is used as a reliable phrase for article determination.

In this paper, significance level β was set to 1%. We can get $\chi_\beta^2 = \chi_{0.01}^2 = 9.21$ from a χ^2 distribution numerical table by assigning a freedom value = 2. In the above example, $\chi^2(C) = 172.88 > \chi_{0.01}^2 = 9.21$, the null hypothesis is rejected, and "covers ϕ wheat" becomes a phrase for article determination.

Here, in the conventional method, reliable phrase (using phrase) condition equal $f(A|C) > 10$.

3.2 Proposed Method using WordNet Thesaurus

To solve the data sparseness problem, we propose a new method that creates new phrases for article determination ("extended phrases") from the phrases of the conventional method ("conventional phrases") using the WordNet thesaurus. Here, we use hypernyms in the WordNet thesaurus. For example, the hypernyms of *wheat* are *cereal*, *grain* and *yellow*. Figure 1 shows an example of the data structure of a WordNet hypernym (hypernyms of *wheat*). Here, "[.....#n#...]" in Figure 1 denotes a hypernym.

Section 3.2.1 describes the procedure of our proposed method.

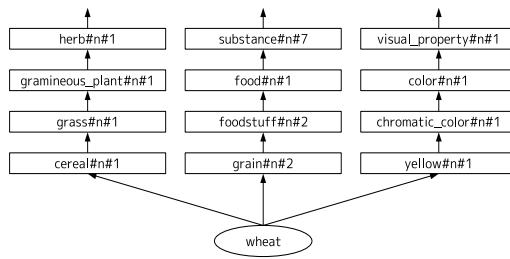


Figure 1: Example of data structure of WordNet hypernyms

3.2.1 Procedure of Proposed Method using WordNet Thesaurus

The procedure of the proposed method has four steps (overall procedure is described in Fig. 2). In the first step, the hypernyms of all nouns in the conventional phrases for article determination are extracted from WordNet. Extended phrases for article determination are created by replacing the nouns in the conventional phrases with their hypernyms. If one conventional phrase has two or more nouns, replacement with the hypernym occurs only one at a time. In other words, N extended phrases for article determination are created by a conventional phrase having N nouns. For example, the following conventional phrase for article determination (conventional phrase; frequency): would create a set of extended

ϕ further software testing ; 30

phrases for article determination (extended phrase; frequency): In the second step, all extended phrases

ϕ further [code#n#3] testing ; 30
 ϕ further software [investigation#n#2] ; 30

for article determination are sorted alphabetically. If the same extended phrases are created from different conventional phrases, the same phrases are sequentially repeated as follows: In the third step, the same extended phrases are merged into one and their frequencies are counted. For example, the extended

ϕ	further	[code#n#3]	testing	; 30
ϕ	further	[code#n#3]	testing	; 90
ϕ	further	[code#n#3]	testing	; 15
ϕ	further	software	[investigation#n#2]	; 30
ϕ	further	software	[investigation#n#2]	; 25
⋮				

phrases for article determination in the above example give: In the fourth and final step, the conditional

ϕ	further	[code#n#3]	testing	; 135
ϕ	further	software	[investigation#n#2]	; 55
⋮				

probabilities are estimated and the highest probability phrases in the conventional phrases and the extended phrases for article determination is used, as in the conventional method.

3.2.2 Words with Multiple Meanings

Many English nouns are multisense words: they have multiple meanings. One multisense word usually has multiple hypernyms. For example, if a multisense word has three hypernyms like wheat in Fig. 1, a replacement with a hypernym occurs three times. In other words, three extended phrases for article determination are created by a conventional phrase having only one noun that has three hypernyms. However, we watch three extended phrases carefully, some phrases often are not appropriate. For example, if we have a conventional phrases (conventional phrase; frequency):

eat ϕ chicken ; 30 ,

and create a set of extended phrases (extended phrase; frequency): the latter extended phrase is not

eat	ϕ	[poultry#n#2]	; 30
eat	ϕ	[domestic_fowl#n#1]	; 30

appropriate. Because [domestic_fowl#n#1] is countable, no article usage is ungrammatical. Therefore, without selecting a relevant hypernym, the system performance remain constant. Consequently, in the proposed method, we must select a relevant hypernym when we replace a multisense word with its hypernym.

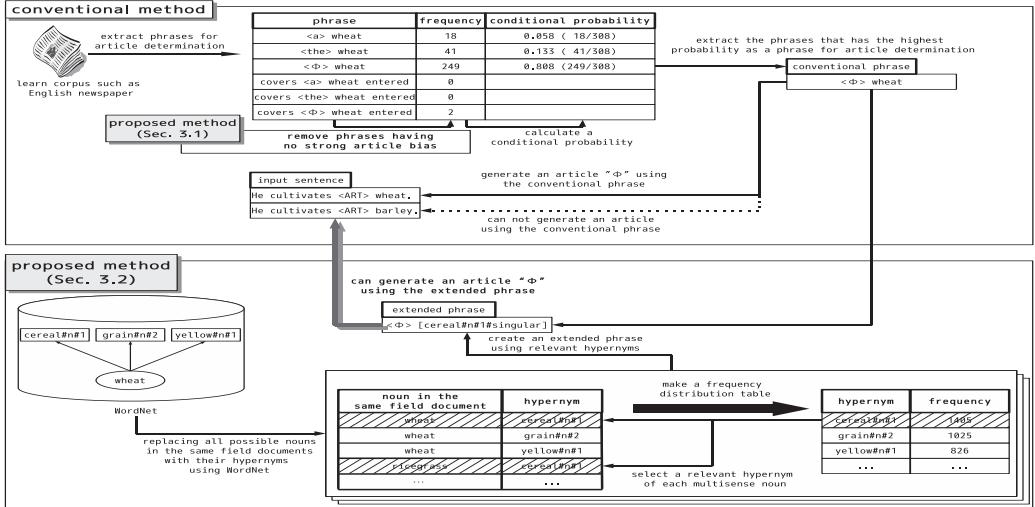


Figure 2: Overall procedure

Selecting relevant hypernyms of multisense words requires understanding of their context, which however, is impossible because the proposed method does not use context understanding but only surface information. Therefore, we use the heuristic that documents in the same field often have nouns with identical hypernyms.

In this paper, we used the topic tags in the Reuters-21578 corpus (Lewis, 1997) as the basis for the same field. In other words, we regard documents with identical topic tags as the same field documents.

As practical procedures, we extract all the hypernyms of all nouns and their frequencies from the corpus and make a frequency distribution table of each field. Here in the proposed method, each multisense word is replaced with a hypernym that has the highest frequency in the frequency distribution table of its document's topic.

For example, multisense word *wheat* has multiple hypernyms ([cereal#n#1], [grain#n#2], and [yellow#n#1]). Here, if a conventional phrase (conventional phrase; frequency): is created from a *grain*

$$\phi \text{ wheat ; } 249$$

topic document and the following is the frequency

distribution table of the *grain* topic: *wheat* is re-

hypernym	frequency
[cereal#n#1]	1405
[time_period#n#1]	1226
[grain#n#2]	1025
[metric_weight_unit#n#1]	956
[yellow#n#1]	826
:	:
:	:

placed with only [cereal#n#1], and the following is the extended phrase created from the conventional phrase (extended phrase; frequency):

$$\phi \text{ [cereal#n#1] ; } 249 .$$

3.2.3 Singular and Plural Words

WordNet does not separate singular and plural noun. If a hypernym: [domestic_fowl#n#1] is selected as the hypernym of chicken (see Section 3.2.2), for example, the following conventional phrase:

$$\text{raise } \phi \text{ chickens}$$

would create a extended phrase:

$$\text{raise } \phi \text{ [domestic_fowl#n#1] .}$$

This rule generate no article for the following input sentence:

raise << Article >> chicken ,

however, *chicken* in this case is countable. Therefore, no article usage is ungrammatical. Consequently, in the proposed method, we must separate hypernyms of singular and plural noun. Specifically, we added a symbol: "#singular" to the hypernyms of singular noun. Similarly, a symbol: "#plural" to the hypernyms of plural noun. In above example, we replace *chicken* with [domestic_foul#n#1#singular] and *chickens* with [domestic_foul#n#1#plural].

4 Experiment

4.1 Experimental Conditions

As the test corpus, a subset of Reuters-21578 was used. Reuters-21578 consists of 21578 documents with topics for each document. In this paper, we used the documents of nine main topics. The statistics on the test corpus is shown in Table 1.

Evaluation procedure has three steps. In the first step, all articles in the evaluation documents are removed. In the second step, article determination for the removed parts in the first step using the proposed or conventional method has done. In the third and final step, three metrics (explained in Section 4.2) were calculated under the condition that removed articles in the first step are correct.

Table 1: Statistics on the test corpus

No. of documents	No. of words	No. of articles
4828	811473	180587

4.2 Evaluation Method

We used recall and precision with 10-fold cross validation to compare the conventional and proposed methods. Threshold θ in Section 2.1 was set by 0.1 (i.e., $\theta=0.35, 0.45, \dots, 0.85, 0.95$). R (Recall) and P (Precision) were defined as follows:

$$R = \frac{\text{No. of correct article determination}}{\text{No. of articles in the set}} \quad (3)$$

$$P = \frac{\text{No. of correct article determination}}{\text{No. of articles determination}} \quad (4)$$

We also defined F (F-measure(Rijsbergen, 1979)) as follows:

$$F = \frac{(b^2 + 1) \times P \times R}{b^2 \times P + R}. \quad (5)$$

Here, parameter b in Eq. (5) represents the relative weights of R and P. We set $b = 1$ in the experiments as the harmonic mean between recall and precision.

4.3 Experimental Results and Discussion

Figures 3, 4, and 5 show the experimental results.

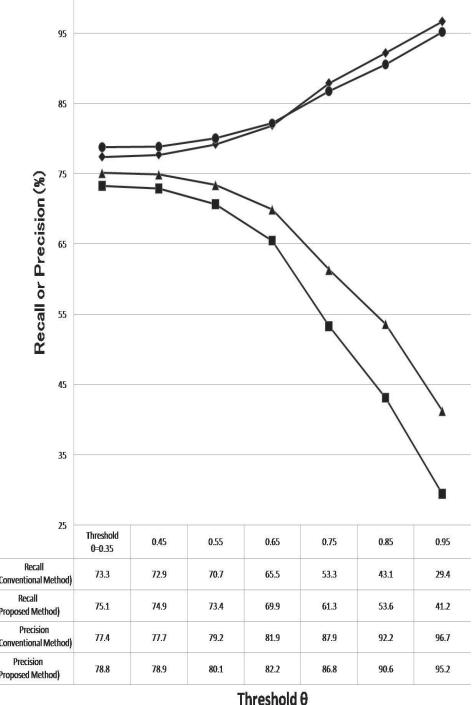


Figure 3: Recall and precision

Figure 3 shows the recall and precision of the conventional and proposed methods. In recall, the proposed method outperformed the conventional method at every θ . This means that new applicable rules are created by the proposed method using the hypernyms of the WordNet thesaurus.

On the other hand, in precision, the proposed method was slightly outperformed by the conventional method at high θ . This may be because the

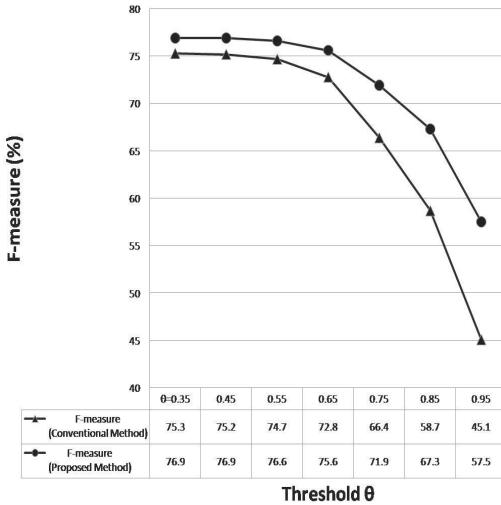


Figure 4: F-measure

ambiguity of the extended phrases for article determination is increased by using the WordNet thesaurus.

Furthermore, in precision, the proposed method outperformed the conventional method at low θ . This may be because the total number of uncorrect article determination decrease by using reliability of phrases for article determination.

Figure 4 shows the F-measure of the conventional and proposed methods. The F-measure of the proposed method outperformed the conventional method at every θ . This means that its precision was slightly outperformed by the conventional method but the system's accuracy improved. Here, F-measure at $\theta = 0.95$, which is high precision but low recall, shows a pronounced 12.4% improvement.

threshold θ	0.35	0.55	0.75	0.95
sub method 1	75.5	75.0	67.8	49.3
sub method 2	76.2	75.8	70.0	52.0

The above table shows the effect of selecting relevant hypernyms described in Sec. 3.2. Sub method 1 shows F-measure of simply using hypernyms of WordNet Thesaurus(Sec.3.2.1). Sub method 2 uses methods described in Sec.3.2.2 and Sec.3.2.3 (and

also in Sec.3.2.1).

Figure 5 shows the precision-recall curve of the conventional and proposed methods. The precision-recall curve plots precision as a function of recall. In Figure 5, the precision-recall curve of the proposed method lies in the upper right corner compared to the conventional method. This means that the total performance of the system is improved.

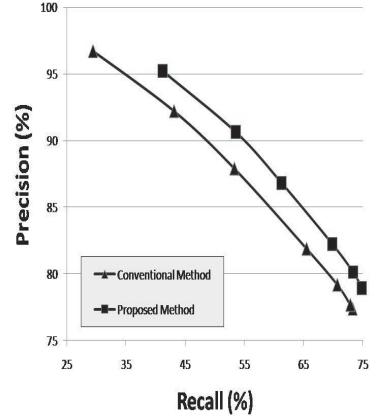


Figure 5: Precision-recall curve

From the F-measure and the precision-recall curve, the proposed method outperformed the conventional method.

5 Conclusions and Discussions

In this paper, we proposed two approach to overcome the data sparseness problem. In the first approach, we use a statistical test to efficiently select and apply reliable phrases for article determination. In the second approach, we propose to expand phrases for article determination by using the hypernyms of the WordNet thesaurus with information about singular/plural noun. Most contributed one is to replace number with [NUM] feature(e.g. 3.75 → [NUM]). In addition, we use heuristic to make an effort to select a relevant hypernym for a multi-sense word. The heuristic proposed in Sec.3.2.1 is not specialized for discriminating count/non-count usage of hypernyms of polysemy. Of course, word sense disambiguation(WSD) is closely related to

the count/non-count usage, the WSD heuristics contributes to generate correct article.

Our research goal is to generate article more precisely by using heuristics that • If words have similar senses, those article usage also become similar • . Therefore, strictly precise WSD is not required, and we think research evaluation has to be done by article generation precision etc., not by WSD. Other WSD topic on count/non-count usage of noun, but not treated in this paper, is semantic feature granuality of WordNet. For example, noun • **chocolate** • has [food],[beverage],[color name] features, but if we success to select [food] feature, but it still have count and non-count usage.

Our experiments showed that the proposed method outperformed the conventional method (Nagata et al., Aug.2005). The proposed method improved recall with only a slight lowering of precision. The F-measure at $\theta = 0.95$, which has the highest precision but the lowest recall point, showed 12.4% improvement. In addition, by using only the statistical test (Sec.3.1), the F-measure at $\theta = 0.95$ showed 6.4% improvement. Similarly, by using only the WordNet thesaurus (Sec.3.2.1), the F-measure at $\theta = 0.95$ showed 6.9% improvement.

Here, the current system only uses the most adjacent level hypernyms of nouns (e.g.; in. Fig. 1, we use [cereal#n#1] / [grain#n#2] / [yellow#n#1] and do not use [grass#n#1] / [foodstuff#n#2] / [chromatic_color#n#1] etc. for the hypernum of wheat). Using higher level hypernyms may improve recall. On the other hand, precision may decrease because the ambiguity of article generation rules increases. Consequently, determining the proper level is very difficult. Determining the proper level is future work.

Acknowledgments

We would like to thank Naoki Isu for his advice on the statistical method. We are also thankful to Keisuke Goto for his technical support.

References

- A. Kawai,K. Sugihara, N. Sugie. 1984. *ASPEC-I : An error detection system for English composition*, Proc. IPSJ Journal (in Japanese), volume 25. Japan.
- E. Izumi, K. Uchimoto, T. Saiga, T. Supnithi, and H. Isahara. 2003. *Automatic error detection in the Japanese learners' English spoken data*', *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*.
- Knight, K. and I. Chander. 1994. *Automated postediting of documents*, In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.
- Lee J. 2004. *Automatic Article Restoration*, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* :31–36.
- Han N. Chodorow M. and Claudia L. 2004. *Detecting errors in English article usage with a maximum entropy classifier trained on a large, diverse corpus*, *Proceedings of the 4th international conference on language resources and evaluation*.
- Minnen G., B. Francis and Copestake A. 2000. *Memory-based learning for article generation*, *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning*.
- R. Nagata, T. Wakana, F. Masui, A. Kawai , and N. Isu Oct.2005. *Detecting article errors based on the mass count distinction*, *Lecture Notes in Artificial Intelligence*. Dale, R., Wong, K.F., Su, J., Kwong, O.Y. (Eds.), Springer-Verlag:815–826
- R. Nagata, T. Iguchi, Y. Furuichi, F. Masui, A. Kawai, N. Isu Aug.2005. *Extracting Collocations for Determining Articles in English Writing*, Proc. PACLING2005.
- F. Bond. 2005. *Translating the Untranslatable: A Solution to the Problem of Generating English Determiners*, CSLI Publications.
- C. Fellbaum. 1998. *WordNet : An Electronic Lexical database*. The MIT Press.
- D. Lewis. 1997. *Reuters-21578 text categorization test collection*.
- C. J. Van Rijsbergen. 1979. *Information Retrieval*, 2nd ed. Butterworths, London.

Linking specific and generalist knowledge

Building terminological resources from sales catalogues and generalist resources

Benoît Trouvilliez^{1,2}

(1) Université Lille - Nord de France, Artois, F-62307 Lens
CRIL, F-62307 Lens / CNRS UMR 8188, F-62307

rue Jean Souvraz SP18, F-62307 Lens France

(2) Onyme SARL, 165 Avenue de Bretagne, 59000 Lille, France
btrouvilliez@onyme.com, benoit.trouvilliez@gmail.com

Abstract

Automatic constitution of specific linguistic resources remains a major challenge in the language processing and terminology domains. The collect of linguistic informations can be manual if human resources and time process are not prohibitive. Besides, this is one of the used approach to constitute stable generalist resources. Specific domains evolve quickly and it becomes necessary to perform automatic or semi-automatic processes. We propose an automatic process to link natural language queries and textual informations in a database (eg a sales catalogue). That uses a generalist language resource for extending knowledge of specific business domain. The process allows to connect the specific vocabulary in the database and the generalist vocabulary in the generalist language resource. This paper describes pre-processing and linking methods. The resulting system is evaluated in a product research context using free text request. This research is based on the linguistic resource built by the process described below.

Introduction

In the early 90s, the researchers in terminological resource construction field worked mainly on domain specific terminological resources (Rastier, 1991). Nowadays, these resources are also tuned to the tasks to be performed. In (Bourigault et al., 2004), terminological resources are built for applications in medical, legal and industrial domains, based on information extracted from specialized text corpora. These examples highlight the need for task-related lexical information. Those information can be extracted from available corpora by some domain expert. Fully automated analysis of the corpora requires some method in order to make so without this manual step of the process. This

challenge can be tackled by successfully linking general lexicon knowledge to task-specific vocabulary. The resulting specific lexicon is obviously less accurate and reliable than an expert built resource, but it is also much less costly to build, easily updated. We also establish in this paper that such lexicon can successfully be used for handling natural language search over some catalogues.

Thus, we focus on the terminological resource construction from sales catalogues. Catalogues contain different products to sell with various attributes (prices, colours, brands). The goal is to link this information to generalist language resources in order to increase the semantic knowledge. That allows the naive natural language research of specific products.

First section is dedicated to the study of free text queries on sales catalogues. In the second section, we present our strategy to build terminological resource. We compare our approach with existing methods and show the expected benefits. The final section is dedicated to the evaluation of our resource and its ability to provide relevant answers to natural language queries. The study is mainly done for French language. But, some results and examples are provided in English.

1 Free text queries on sales catalogues

1.1 Main framework

The work described in this paper is part of an ambitious project, whose goal is to build a conversational agent able to handle clients throughout the whole sale process. The project encompasses social interaction, search refinement and more. This paper focuses on building language resource fit for translating natural language product search into database query. Figure 1 illustrates the process.

1.2 Available resources

We assume that the catalogue database is the only task-related resource available. It has been

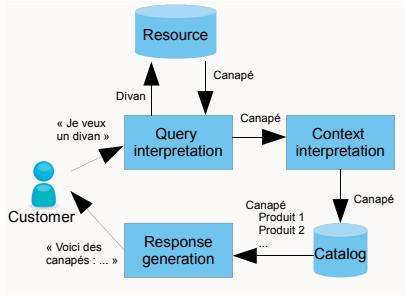


FIGURE 1 – Requests on sales catalogues

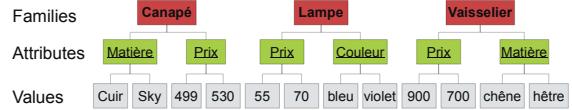


FIGURE 2 – Simplified structure of product catalogue

Yaourtiere
Tringle, barre a rideau
Vidéoprojecteur
Tiroir
Wok, tajine
Tondeuse homme
Vêtement de travail homme
Thermomètre bébé
Chaine hifi

FIGURE 3 – Family labels

Family Label	Attribute Label	Attribute Unit
<i>Chaine hifi</i>	<i>Profondeur (deep)</i>	mm
<i>Chaine hifi</i>	<i>Hauteur (high)</i>	mm
<i>Chaine hifi</i>	<i>Largeur (large)</i>	mm
<i>Chaine hifi</i>	<i>Prix (price)</i>	€
<i>Chaine hifi</i>	Brand	
<i>Chaine hifi</i>	Wifi	
<i>Chaine hifi</i>	<i>Enregistreur MP3 (MP3 recorder)</i>	
<i>Chaine hifi</i>	HDMI	
<i>Chaine hifi</i>	<i>Lecteur/Enregistreur DVD</i> (DVD reader/recorder)	
<i>Chaine hifi</i>	Description	
<i>Chaine hifi</i>	Reference	
<i>Chaine hifi</i>	Watts	

FIGURE 4 – Attributes of family *Chaine hifi*

Attribute Label	<i>Chaine hifi 1</i> values	<i>Chaine hifi 2</i> values	<i>Chaine hifi 3</i> values
Brand	Philips	Philips	Muse
Prix (price)	129.00	168.00	128.00
Reference	<i>Chaine Hifi DVD USB</i> <i>Philips MCD122</i>	<i>Chaine Hifi DVD USB</i> <i>Philips MCD712</i>	<i>Microchaine DVD MP3</i> <i>Muse M-68DM</i>
Description	<i>Lecture via USB direct et une expérience sonore 2x20W</i>	<i>Intisifiez votre expérience sonore avec un son 2x50W et une lecture USB directe</i>	<i>Une qualité d'écoute de 2x20W dans un design épuré</i>

TABLE 1 – Attribute values for *Chaine hifi* family products

built by a semi-automatic process and will provide a basic taxonomy for the constitution of domain-specific vocabulary. This catalogue replaces domain-specific text corpus of vocabularies and specific resources, as used in some terminology building approach (Bourigault et al., 2004). In this taxonomy, products are organized in families such as sofas, lamps or shirts through several textual fields (Table 3). Those families are biased toward the sale task, i.e. products in one family are likely to be found on the same department store. Those products also share a set of attributes (Table 4). Each product may have some value defined for any attribute in its family (Table 1). Those values are products of some semi-automatic process applied to raw data in order to build the catalogue. Therefore, informations contain in this fields can be wrong or missing. Figure 2 illustrates the taxonomy.

1.3 Problems and propositions

The catalogue has been designed for sales tasks, and is not fit even for direct full text search. For example, it is of no use if searching a product with anything but the exact description, or part thereof.

We see in section 3.1 how to extract a useful lexicon from catalogues, thus overcoming some limits and shortcomings of this most important resource.

However, some problems can not be handled by relying only on the catalogue. For example, the customer requests *serviettes* yield no answer, although the catalogue features *draps de bain*. This example illustrates synonymy-related problems. More, the catalogue appears unable to generalize or specialize families and attributes of the products. So no answers are returned if the customer requests a cupboard despite the catalogue contains dressers. It is all the same if the catalogue includes a guitar family and a ukulele is requested. Such problems are to be excepted as customers have no way to know neither the exact types of products sold nor the label in the database. This part of the task would obviously require the acquisition of pragmatic knowledge in order to understand the user's request. So, we need to enrich extracted knowledge from the catalogue with generalist resources to better support free text queries. We present these resources in the next section.

2 Generalist resources

General and freely available knowledge resources are accessible on the Internet. Wikipedia

has so been used to acquire knowledge on the most common French mistakes in (Wisniewski et al., 2010) and to transcribe an English language resource automatically in French in (Sagot and Fišer, 2008). Such resources could be used to extract pragmatic knowledge but it remains a difficult challenge due to the complexity of the extraction.

Resources built semi manually by linguists to describe languages can also be used. The most widely used and acknowledged is the Princeton Wordnet (Miller, 1995), created for the English language. It is the base of Wordnet resources with similar characteristics, developed for more than 70 languages and listed by the Global Wordnet Association¹. The centerpiece is the synset, i.e. a sense unit, or a language concept. Synsets are themselves organized into a graph where links depict generalization (hyperonymy), specialization (hyponymy), part/whole relation (meronymy)

The latter approach seems the better for our work due to the existing language concept structure that can be used to determine the proximity between the search and the answer thereof. The proximity found is more linguistic than pragmatic but we suppose that this proximity can help us to solve the problems shown previously. We process with the Free French Wordnet (WOLF) (Sagot and Fišer, 2008) and EuroWordnet (Vossen, 1998) for French language and with the EuroWordnet (Vossen, 1998) for English one. We argue this choice more precisely in section 3.2.

3 Integrating generalist knowledge

Links should be established between catalogue's knowledge and generalist resources' knowledge. Figure 5 illustrates our process described below.

3.1 Building specific lexicon

We need to extract the most relevant informations for the search interpretation. We have identified three types of fields :

1. family and attribute labels (sofa, lamp,...) (material, price,...)
2. attribute units (€, cm,...)
3. different values for each product attribute (leather, oak,...)

¹. reference : <http://www.globalwordnet.org> (accessed 08/12/2011)

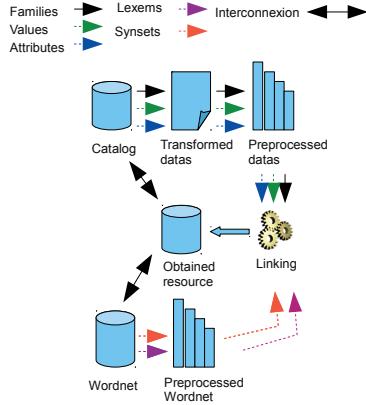


FIGURE 5 – Linking process diagram

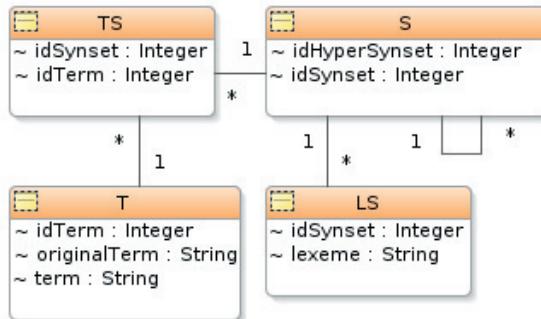


FIGURE 6 – Relational model

Selected informations must be cleaned and pre-processed as they are non adapted to the task. Informations thereof constitute a specific lexicon.

3.1.1 Specific taxonomy adaptation

A generalist taxonomy would have a family for every type of clothing (eg : shirts, pants,...) and an gender attribute specifying the genders available for this type of clothing (eg : man, woman, child,...). But, marketing taxonomies in catalogues have a family for each type and kind of clothing (eg : man shirt, woman shirt,...) to suggest an aspect of shelves. These aspects make non-canonical relations between the two lexicons. To overcome this problem, the inadequate presence of words such as man, woman, child in family labels must be managed. A gender attribute is so added to the families with this characteristic and completed with the appropriate value for each product. Table 2 illustrates some examples of filtering on family labels.

Similarly, for marketing purposes, family labels contain lists of comma-separated concepts. The T-shirt, jacket family is an example. We hypothesize that it would be interesting to transform those concept associations in catalogues to synonymy links in the final resource. It should be noted that concept associations are not necessarily inconsistent with the synonymy in generalist resources as is the case with the Television, Tv family. The exploitation of these differences between synonyms and taxonomic concept associations is studied more precisely in the section dedicated to the linking process in itself. To build this new synonymy links, different concepts must be isolated from each other. Table 2 illustrates some English and French examples. However, in some

examples like the pillow case, bolster case family, our segmentation process gives labels sometimes incomplete due to phenomena of elision of terms recurring in the enumerative list. It is elided in pillow, bolster case to avoid repetition of the word case. As a result, the segmentation process fails for pillow case. For French, it should be possible to achieve distributive on labels containing words such as *de* or *à*.

3.1.2 Spell checking

We handle data obtained from reliable sources in terms of spelling as it is marketing data and language resources widely available. However, fields can contain uppercase while they are common names (eg : Talkie Walkie) or else not have any accent (eg : Television instead of *Télévision* in French). A spelling correction seems necessary. Nevertheless, we use an insensitivity of our procedure to accents and case due to both the quality of the resources used and the relatively strong constraints in terms of processing cost (CPU time and memory used). This solution limits our ability to correct.

3.1.3 Lemmatization / Stemmatization

Some values of the knowledge database are filled in plural while linguistic informations are filled in singular. To overcome this problem, it is usual to lemmatize the forms of the same word in its canonical one. The TreeTagger tool (Schmid, 1994) can perform that automatically. Nevertheless, we use instead a simplified stemmatization process due to the quality of resources used and the time constraints. All words ending with 's' or 'x' are considered in plural and correspondences

Family labels from a catalogue	Cleaned labels	Terms found			Simplified stemmatized terms		
Drap plat sheet	Drap plat sheet	Drap plat sheet			Drap plat sheet		
Manteau, veste femme woman coat, jacket	Manteau, veste coat, jacket	Manteau	veste		Manteau	veste	
Manteau, veste sport femme woman sport coat	Manteau, veste sport sport coat	Manteau	veste sport		Manteau	veste sport	
Tee-shirt, polo, chemise tee-shirt, polo, shirt	Tee-shirt, polo, chemise tee-shirt, polo, shirt	Tee-shirt	polo	chemise	Tee-shirt	polo	chemise
Television, Tv	Television, Tv	Television	Tv		Television	Tv	
Tringle, barre a rideau curtain rod	Tringle, barre a rideau curtain rod	Tringle	barre a rideau		Tringle	barre a rideau	
Taie d'oreiller, de traversin pillow, bolster case	Taie d'oreiller, de traversin pillow, bolster case	Taie d'oreiller	de traversin		Taie d'oreiller	de traversin	
Tapis carpet	Tapis carpet	Tapis	carpet		Tapis	Tapi	
Tapis de bain shower mat	Tapis de bain shower mat	Tapis de bain	shower mat		Tapis de bain	Tapi de bain	
Chaussures sport sports shoes	Chaussures sport sports shoes	Chaussures sport	sports shoes		Chaussures sport	Chaussure sport	

TABLE 2 – English and French catalogue data adaptation and preprocessing

should be found *both* with the *original terms* and those *without this final symbol*. This solution limits our ability to lemmatize. With the French term *chevaux*, we attempt to link with *chevau* instead of *cheval*. However, the strategy to keep in any case the original term allows to guarantee an absence of degradation of the overall process by the addition of this preprocess in the limit where the term proposed as lemma is not an existing term unrelated to the original one. Table 2 presents some results of this preprocess.

Subsequently, text fields obtained in the specific lexicon are denoted by terms.

3.2 Building generalist lexicon

Second part of the task concerns generalist resources. We need to extract the most relevant informations for the task. So, a generalist lexicon is built containing only interesting parts of the resources for the task.

3.2.1 Lexemes

First relevant entities are lexemes. Preliminary studies had shown that almost only nominal lexemes are relevant for specific data. Only them are kept. That allows in the specific case of sales catalogues to simplify the procedure by eliminat-

ing irrelevant lexemes. Its relevance is therefore to study as the specific data type to handle. More, if the nominal sewing machine lexeme is not present in the language resource, this filter prevent from finding a link via the sewing lexeme. Non nominal lexemes could be used to overcome a lack of nominal lexemes. Both of resources studied in section 2 are relevant as they contain lexemes.

3.2.2 Linguistic link types

Second relevant entities are linguistic link types. As we show in section 1.3, we need informations about synonymy and generalization of concepts. Wordnet resources have many types of links between different concepts of language like those we need. So, only synonymy and hyper/hyponymy links are kept. Again, the relevance of this limitation is to study depending on the wished process below. Wordnet seems better as it provides those needed links « out of the box » contrary to resources like Wikipedia. Other links like mero/holonymy could be interesting.

3.3 Integrating generalist lexicon

3.3.1 Relational representation

Let be T , a relation representative of the specific lexicon. Let be S and LS , relations repre-

senting synsets, lexical items (lexemes), links between synsets and links between lexemes and synsets in the generalist lexicon. Relation TS is filled by the process described below. The data structure of these relations is given in figure 6. Table T is supplied for $idTerm$ and $originalTerm$ by the corresponding original catalogue fields while the field $term$ is one of the terms obtained at the output of the preprocessing described in section 3.1 for the field $originalTerm$ in input. As many lines as necessary are inserted for each entry.

3.3.2 Integrating synonymy

First, synonymy links are integrated. Strict equivalences are searched between specific and generalist lexicons. Relation TS is filled (Equation 1).

3.3.3 Integrating hyperonymy

Second, hyperonymy links are integrated. Hyponymies are searched in the specific lexicon relative to generalist one. In the examples sport shoes and shoes, the first one contains strictly more information than the latter. We assume that a label which contains strictly more information than another is an hyponym of the second. Two separate cases depending on generalist lexeme are considered :

- it is simple, ie not composed
- it is composed of two words as less

Specific data selection for hyperonymy integration Hyperonymy integration is required for entries in specific lexicon that are not related by synonymy. Relation NLT is filled with those entries of relation T .

Integrating hyperonymy for non composed lexemes Relation NLT contain entries for which one of the words constituting the terms is known as linguistic data. These inputs and their linguistic corresponding are stored in the relation $BTLS$ (Equations 2).

Integrating hyperonymy for composed lexemes Relation NLT contains entries for which the original sequence of characters, corresponding to the term, includes strictly a sequence of characters corresponding to a linguistic data. These inputs and their linguistic corresponding are stored in the relation $STLS$ (Equation 3). This is relevant only by its exclusive application to composed lexemes. Otherwise, some wrong associations appear by the possibility of identifying the shortest lexemes in many other words. French lexeme « *vie* » could be

associated with « *éVIEr* » in this part of the procedure without that limitation.

Linguistic relations completion Linguistic relations are filled using the two subsets identified in the previous hyperonymy integration steps (Equations 4).

3.3.4 Enrichment of lexemes

Informations connected by synonymy and hyponymy are used to enrich lexemes. Relation T contains terms that do not correspond to a label present in the linguistic resources. Nevertheless, a link was found between the specific knowledge and the linguistic resources (relation TS). Relation LS is filled with those informations (Equations 5).

4 Evaluation of the obtained resource

4.1 Generalities

Execution environment requires to have the smallest possible process time (less than 10 minutes is ideal). We postulate that a full implementation of the process on families and a partial one on attributes, units and values can ensure this processing time while retaining interesting results. It includes only synonymy integration for attributes, units and values. Two environments are considered : a time restricted one is denoted by restricted mode while an other non restricted is denoted by normal mode. Evaluations presented below compare results in them.

4.2 Statistical evaluation

A statistical evaluation of the process is performed on four French sales catalogues having various thematics and volumes to measure its stability and its effectiveness in producing a quality resource : a generalist catalogue with clothing, furniture, accessories, computers, appliances ... and three specialized catalogues in wine, high-tech and car products. It is also performed on an English sales catalogue specialized in sport equipments. Because this database features neither attributes nor units, results are only computed on families and values. Results are given in Table 3 and in normal mode by default. Results in restricted mode, if different, are given in parentheses. But this restricted strategy can not find links between the linguistic resources and the most complex labels and values. The number of attributes and values considered for the evaluation is so limited.

$$TS \leftarrow \Pi_{idTerm, idSynset}(T \bowtie_{lexeme=term} LS) \quad (1)$$

$$\begin{aligned} Words &\leftarrow \{(idTerm, sterm) \mid \exists(idTerm', term) \in NLT, idTerm = idTerm' \wedge sterm \in split(term)\} \\ WordsTLS &\leftarrow \Pi_{idTerm, idSynset}(Words \bowtie_{lexeme=term} LS) \end{aligned} \quad (2)$$

$$SequenceTLS \leftarrow \{(idTerm, idSynset) \in NLT \times LS \mid \exists lexeme \in LS, seq(lexeme) \subset seq(originalTerm)\} \quad (3)$$

$$\begin{aligned} TLS &\leftarrow WordsTLS \cup SequenceTLS \\ TLSUID &\leftarrow \rho_{idTerm, idHyperSynset, idSynset}\{(idTerm, idSynset, idSynset') \in TLS \times \mathbb{N} \mid \exists! idSynset' \in \mathbb{N},\right. \\ &\quad \left.idSynset' = uid(idSynset, idTerm)\right\} \\ S &\leftarrow S \cup \Pi_{idHyperSynset, idSynset}(TLSUID) \\ TS &\leftarrow TS \cup \Pi_{idSynset, idTerm}(TLSUID) \end{aligned} \quad (4)$$

$$\begin{aligned} TLS &\leftarrow T \bowtie_{(originalTerm=lexeme) \vee (term=lexeme)} LS \\ NewLTS &\leftarrow \rho_{idTerm, newLabel}((\Pi_{idTerm, originalTerm}(T) \cup \Pi_{idTerm, term}(T)) - \Pi_{idTerm, lexeme}(TLS)) \bowtie TS \\ LS &\leftarrow LS \cup \Pi_{idSynset, newLabel}(NewLTS) \end{aligned} \quad (5)$$

with :

- $split(a)$ which cuts a string as the space character.
- $seq(a)$ which returns an ordered set of characters in the string according to their order of appearance.
- $uid(a)$ which returns an unique identifier.

		generalist catalog			wine catalog			high-tech catalog			car catalog			sport catalog
		Wolf (W)	EuroWord-net (EWN)	W+EWN	W	EWN	W+EWN	W	EWN	W+EWN	W	EWN	W+EWN	EWN
	Language	French			French			French			French			English
Families	#	273			7			92			6			52
	# linked	198	226	237	4	4	4	66	68	75	3	3	4	43
	% linked	72.5%	82.8%	86.9%	57.1%	57.1%	57.1%	71.7%	73.9%	81.5%	50%	50%	66.7%	82.7%
Attributes	#	12			16			43			56			-
	# linked (partial)	9 (7)	10 (8)	10 (8)	9 (7)	7 (5)	9 (7)	31 (20)	31 (20)	31 (20)	42 (12)	43 (13)	44 (13)	-
	% linked (partial)	75% (58.3%)	83.3% (66.7%)	83.3% (66.7%)	56.3% (43.8%)	43.8% (31.3%)	56.3% (43.8%)	72.1% (46.5%)	72.1% (46.5%)	72.1% (46.5%)	75% (21.4%)	76.8% (23.2%)	78.6% (23.2%)	-
Values	# (partial)	21120 (3792)			8728 (360)			3612 (3496)			26337 (1420)			5475 (645)
	# linked (partial)	13469 (232)	13743 (237)	16296 (331)	2686 (17)	4018 (13)	4462 (22)	565 (214)	414 (129)	674 (259)	2976 (127)	2889 (121)	3662 (180)	3767 (111)
	% linked (partial)	63.8% (6.1%)	65.1% (6.3%)	77.2% (8.7%)	30.8% (4.7%)	46% (3.6%)	51.1% (6.1%)	15.6% (6.1%)	11.5% (3.7%)	18.7% (7.4%)	11.3% (8.9%)	11% (8.5%)	13.9% (12.7%)	68.8% (17.2%)
Units	#	63			3			46			21			-
	# linked (partial)	27 (5)	21 (2)	30 (6)	3 (1)	2 (0)	3 (1)	22 (10)	17 (5)	23 (11)	10 (8)	2 (1)	11 (9)	-
	% linked (partial)	42.9% (8.0%)	33.3% (3.2%)	47.6% (9.5%)	100% (33.3%)	66.7% (0%)	100% (33.3%)	47.8% (21.7%)	37% (10.9%)	50% (23.9%)	47.6% (38.1%)	9.5% (4.8%)	52.4% (42.9%)	-
	Time in mn (partial)	-	-	1046.87 (7.61)	-	-	627.95 (1.15)	-	-	929.05 (5.78)	-	-	333.93 (2)	333.28 (1.36)

TABLE 3 – Statistical results of our linking method over French and English processes

Corpus	generalist catalog	wine catalog	high-tech catalog	car catalog	sport catalog
Language	French	French	French	French	English
# messages	69	20	18	20	20
F-measure on families (partial)	52% (50%)	42% (48%)	58% (62%)	32%	72% (78%)
F-measure on attributes (partial)	50% (49%)	45% (49%)	65% (70%)	30%	57% (62%)
% expected answers (partial)	36% (38%)	20% (15%)	61% (67%)	15%	70% (65%)
Time for all messages in mn (partial)	2.21 (1.6)	1.03 (2.16)	1.33 (1.35)	0.15	0.31 (0.35)
Average time for a message in s (partial)	1.9 (1.4)	3.1 (6.5)	4.4 (4.5)	0.5	0.9 (1.1)

TABLE 4 – Results in product research

The best statistical French results (in bold in the table) are always at least obtained by the resource built from Wolf and EuroWordnet. In a majority of cases, it even manages to be strictly better than those obtained from each of the two. The generalist catalogue linking is quite better than the three specialized catalogues linking in particular for values as those entities are more specialized in the latter catalogues. Another interesting point concerns the lower quality of the characteristic, value and unit links in restricted mode as compared to normal mode. But as expected, the process is in average 250 times faster in restricted mode as compared to normal mode. The results recorded in English are quite similar to the best French cases with a linking rate on the families of the order of 80% and a linking rate on the values of the order of 70% in normal mode despite the specialization of the catalogue in sport equipments values. In restricted mode, there is even a slight increase on the values with a success rate of the order of 20% against an order of 10% in the best French case.

4.3 Product research process evaluation

The final goal of this resource is to provide semantic informations in order to process natural language queries over the catalogue's products. We formed five corpora of texts consisting of customer requests in all five catalogues described above. Four of those five corpora are written in French while the later is in English. Each corpora is then labelled manually according to the expected results. Finally, we evaluate the results returned by the search engine in normal and restricted modes according to three criteria : an F-measure on the families, an F-measure on the attributes and the percentage of results returned same as expected. Table 4 shows the results. English catalogue has a score of F-measure on the attributes identified although it has none in English. This score is computed using requested values during search. The evaluation is done with the resource built from Wolf and EuroWordnet for French and from EuroWordnet for English.

The results show that our system understands correctly almost 30% of the queries for French. Those also show that the restricted mode (results in parentheses) obtained each time a similar score to that obtained in normal mode. The maximum deviation obtained is 7%. The gap between the two modes in the previous evaluation about attributes and values seems to have a limited impact

on the response of the system. The performance of our system appears to be higher in English. It returns 70% of correct answers on the test catalogue. This result is 10% higher than even the best French score.

5 Conclusion

This process allows to connect specific knowledge and generalist resources like Wordnet. Relations are established between the two types of resources by integrating the generalist lexicon into the specific one. Adding linguistic resources in the process tends to improve the statistical linking results. In order to save on processor time, it is possible to restrict the process without restrict final result quality. These results tend to confirm the suitability of our procedure to other languages than French and to confirm that the quality of the resource obtained is highly dependent on the base resource. This sustains the idea that English resources are better rounded than their French counterparts.

Acknowledgements

I thank Pierre Marquis and Vincent Dubois, PhD advisors, and Antoine Serniclay, Thibaud Vibes and Onyme SARL employees.

References

- [Bourigault et al.2004] D. Bourigault, N. Aussenac-Gilles, and J. Charlet. 2004. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, 18(4) :24.
- [Miller1995] G.A. Miller. 1995. Wordnet : a lexical database for english. *Communications of the ACM*, 38(11) :41.
- [Rastier1991] F. Rastier. 1991. *Sémantique et recherches cognitives*. Presses universitaires de France.
- [Sagot and Fišer2008] B. Sagot and D. Fišer. 2008. Building a free french wordnet from multilingual resources. In *Ontolex 2008*.
- [Schmid1994] H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, volume 12. Manchester, UK.
- [Vossen1998] P. Vossen. 1998. Eurowordnet a multilingual database with lexical semantic networks. *Computational Linguistics*, 25(4).
- [Wisniewski et al.2010] G. Wisniewski, A. Max, and F. Yvon. 2010. Recueil et analyse d'un corpus éco- logique de corrections orthographiques extrait des révisions de wikipédia. In *TALN 2010*, Montréal, July. ATALA, RALI et POLYMTL.

A Linguistic Approach to Feature Extraction Based on a Lexical Database of the Properties of Adjectives and Adverbs

Franco Tuveri, Manuela Angioni

CRS4, Center of Advanced Studies, Research and Development in Sardinia,
Parco Scientifico e Tecnologico, Ed. 1, 09010 Pula (CA), Italy.
tuveri@crs4.it, angioni@crs4.it

Abstract

Reviews are used every day by common people or by companies who need to make decisions. A such amount of social data can be used to analyze the present and to predict the near future needs or the probable changes. Mining the opinions and the comments is a way to extract knowledge by previous experiences and by the received feedback. In this paper we propose an automatic approach to the extraction of feature terms by means of the semantic analysis of textual resources. We enriched the contents of WordNet, related to the meanings expressed by adjectives and adverbs, with a set of properties having a positive, negative or objective orientation associated and other properties that could add particular and important information in semantic analysis.

1 Introduction

The pervasive diffusion of social networks as common way to communicate and share information is becoming a valuable resource for analysts and decision makers. Reviews are used every day by common people or by companies who need to make decisions. A such amount of social data can be used to analyze the present and to predict the near future needs or the probable changes. Mining the opinions and the comments is a way to extract knowledge by previous experiences and by the received feedback.

As asserted by Lee et al. (2008), “Opinion Mining can be roughly divided into three major tasks of: development of linguistic resources, sentiment classification, and opinion extraction and summarization”.

A relevant goal of our work is the development of an automatic Opinion Mining system, based on a linguistic approach, operating in a general and not clearly defined domain and able to extract and contextualize features related to products or services.

The term feature is here used with the same sense given by Ding et al. (2008) in their approach to the Opinion Mining.

It seems of considerable interest to explore the issue of the contextualization of features from a set of reviews in order to more easily combine the qualitative information expressed by the users to the highlighted features. Some approaches propose a methodology (Rentoumi et al., 2009) for assign a polarity to word senses applying a Word Sense Disambiguation (WSD) process.

As in Benamara et al. (2007), we propose a linguistic approach to Opinion Mining and, more in details, to the automatic extraction of feature terms by means of a semantic analysis of textual resources. We enriched the contents of WordNet, related to the meanings expressed by adjectives and adverbs, with a set of properties having a positive, negative or objective value associated and other properties that could add particular and important information in semantic analysis.

We focus on the analysis of the opinions through the processing of textual resources, the information extraction and the evaluation of a semantic orientation. The approach performs a semantic disambiguation and a categorization phase and takes into account the meanings expressed in conversations, considering for instance the synonyms of relevant terms.

Working on Opinion Mining with WordNet, we soon realized the needs to have additional resources associated to the qualities of adjectives and adverbs. A set of properties related both to adjectives and to adverbs have been identified and associated to each synset, providing as result a set of thematic categories. The supplementary information given by the properties helps to better identify the opinion referred to features, giving a more detailed description of the features.

The remainder of the paper is organized as follows: Section 2 refers to related works. Section 3 introduces our approach and examines the work performed on the structures of adjectives and

adverbs, giving some details about the semantic classification and the contextualization of features by means of the properties of adjectives and adverbs. Finally, Section 4 explains our approach to future works and draws conclusions.

2 Related Works

Many approaches to Opinion Mining and Sentiment Analysis are based on linguistic resources, lexicons or lists of words, used to express sentiments or opinions. The lack of suitable and/or available resources is one of the main problems in the Opinion Mining process and in general in the analysis of textual resources by Natural Language Processing techniques. Knowing the polarity of words and their disambiguated meanings can surely help to better identify the opinions related to specific features.

SentiWordNet (Esuli and Sebastiani, 2007; Baccianella et al., 2010) is one of the publicly available lexical resources, that extends WordNet thanks to a semi-automatic acquisition of the polarity of WordNet terms, evaluating each synsets according to positive, negative and objective values. It provides the possibility to accept user feedback on the values assigned to synsets, allowing to build a community of SentiWordNet users in order to improve SentiWordNet. Despite its wide coverage SentiWordNet does not give some additional information necessary to contextualize the content of the sentences.

Another lexical resource consisting of WordNet senses automatically annotated by positive and negative polarity, is Q-WordNet (Agerri and García-Serrano, 2010) that tries to maximize the linguistic information contained in WordNet, taking advantage of the human effort given by lexicographers and annotators.

Another resource, available for free for non-profit institution, is WordNet-Affect (Valitutti et al., 2004). It was developed starting from WordNet, assigning one or more affective labels (*a-labels*) to a subset of synsets representing affective concepts that contribute to precise the affective meaning. For example, the *a-label* EMOTION represents the affective concepts related to emotional state. Other concepts are not emotional-affective but represent moods, situations eliciting emotions, or emotional responses.

3 Our Approach to the Feature Extraction

In order to perform a feature extraction system based on linguistic resources, an extension of the

properties of a set of adjectives and adverbs contained in the WordNet has been defined. The resource is used in the sentiment classification and in the opinion summarization tasks.

The first step of feature extraction, a task of the opinion extraction, is the identification of the topic for the corpus. A Semantic Classifier (Angioni et al., 2008) performs the categorization of the corpus defining a set of top categories. The Semantic Classifier is capable to categorize text documents automatically, applying a classification algorithm based on the Dewey Decimal Classification system, as proposed in WordNet Domains (Magnini et al., 2002), a lexical resource representing domain associations among the word-senses of WordNet.

The categories extracted by the classifier define the context, the topic, for the corpus of reviews. The Semantic Classifier parses and categorizes each sentence in order to decide which specific phrases are relevant to the topic.

For example, analyzing reviews about tourism and especially reviews linked to hotels, we expect to examine sentences containing opinions about geographical locations, buildings, rooms, staff, and food.

Syntactic and semantic analyses have been performed for the pos-tagging and the sentence analysis. The identification of adjectives and adverbs and the extraction of information by means of adjectives qualities and a subjective lexical database have a relevant role in this phase.

The definition of a context for the set of reviews is a necessary support even in a next phase when a WSD for the adjectives and adverbs is performed.

As described in next parts of the paper, we define a completely automatic method for the feature extraction from sentences in a general domain. Further steps could be briefly described as follows: distinction between subjective and objective or factual sentences by means of semantic analysis; analysis and validation of the features extracted. The result of this kind of analysis could be used in a summarization step in order to define the polarization of reviews related to the features extracted.

3.1 Properties of Adjectives and Adverbs

We found that some additional properties are necessary in the analysis of sentences in order to improve the performances of the WSD phase or in order to build better meaning relations between nouns and adjectives in particular related to opinions. A lexical database of synsets has

been produced enriching adjectives and adverbs' synsets with a set of properties related both to the emotional/moody sphere, but even to the geographical, behavioral or physical related sphere. The properties provide the possibility to distinguish between subjective, objective and factual having polarity valence sentences. The addition of information helps to better identify the sentiment expressed in relation to the features and the result of the analysis certainly benefits from it.

Some linguistic resources are built considering three properties: subjectivity, orientation, and strength of term attitude. For example, 'good', 'excellent', and 'best' are positive terms while 'bad', 'wrong', and 'worst' are negative terms. 'Vertical', 'yellow', and 'liquid' are objective terms. 'Best' and 'worst' are more intense than 'good' and 'bad' (Lee et al., 2008).

The presented linguistic resource does not contain information about the strength property because it is possible, i.e., to retrieve the basic form of superlative and comparative adjectives. We instead mainly concentrate on the qualitative adjectives able to specify for instance color, size, smell, making the meanings of sentences clearer or more exact. We have thus extended the properties of the semantic network of WordNet focusing on the characteristics of adjectives and adverbs starting from the data retrieved by FreeLing (Atserias et al., 2006; Vossen, 1998) for English, Catalan and Spanish. Once identified a representative number of properties, about 2.300 pairs of adjectives/synsets and about 480 pairs of adverbs/synsets have been manually classified according to a set of attributes identified by their association with nouns and verbs and chosen on the basis of their frequency of use in the language. The work, valid for the english language, has been performed by two evaluators following some predefined rules in order to define the criteria before associating synsets, properties and polarity. The reason of this approach is due to the similarity of some categories like Emotion, Moral/Ethic, and Character, depicted in Table 1, where there is a very slight distinction between their meanings.

A first set of 150 adjectives and adverbs has been selected and evaluated together by the two evaluators in order to align the evaluation criteria. Then, independently, each evaluator performed all adjectives and adverbs evaluation.

Finally the results obtained by the evaluators have been compared in order to emphasize the points of disagreement between them. Every time the categorization by means of the gloss

definition of the synsets or the assignment of the polarity generated discrepancy in the interpretation, the results have been compared and discussed by the two people in order to establish a common evaluation. If a convergence of opinions was not possible, the synset was discarded.

Adjectives	Pos.	Neg.	Obj.	Tot.
Emotion	52	73	3	128
Moral/Ethic	45	155	2	202
Character	355	584	220	1159
Weather	7	25	6	38
Color	0	9	42	51
Quantity	16	0	9	25
Appearance	41	83	46	170
Material	22	11	54	87
Shape	0	0	30	30
Touch	3	13	6	22
Taste	40	41	5	86
Dimension	11	2	60	73
Chronologic	3	0	30	33
Geographic	0	10	19	29
others	29	17	87	133
Total Adjectives	624	1023	619	2266

Table 1: Adjectives' properties

Only the synsets having a complete agreement of both the evaluators in the polarity evaluation and in the association of the properties have been included. The level of disagreement is mainly related to the assignment of synsets to the categories: Moral/Ethic, Character, Emotion. The cause is that sometimes the glosses of WordNet were not so clear in order to decide the most correct interpretation. In this case the reviewers used some dictionaries to find other definitions of meaning of terms. The disagreement is widely affected by the classification of the synset in these three categories, as they represent about the 65% of all the synsets analyzed. Related to the polarity valence of synsets, an agreement near the 100% has been reached because the evaluation of the polarity has been indicated only as positive, negative and objective.

The identified characteristics provide additional information about the content of the sentences, regarding for instance personal, moral, ethical or even aesthetical aspects, as showed in Table 1 where the number of positive, negative and objective adjectives are listed. It is evident that some of these categories allow a polarization that can be used by Opinion Mining algorithms. The last row in Table 1 is related to the number

of adjectives having polarity orientation but not classified in any of the previous categories.

Adverbs	Pos.	Neg.	Obj.	Tot.
Time	0	0	7	7
Manner(things)	18	25	8	51
Manner (persons)	166	205	5	376
Place	0	0	3	3
Intensifiers	0	0	38	38
Quantity	0	0	6	6
AND	1	0	0	1
<i>Total Adverbs</i>	<i>185</i>	<i>230</i>	<i>67</i>	<i>482</i>

Table 2: Adverbs' properties

Adverbs are useful too into the Opinion Mining process. We concentrate on some adverbs classified by their meaning, position or their strength, associating to each of them a specific synset as made for the adjectives.

Adjectives Properties	Examples		
	Pos.	Neg.	Obj.
Emotion	alive	depressed	labial
Moral/Ethic	respectable	caddish	-
Character	audacious	caitiff	vacant
Weather	beautiful	arid	climatic
Color	-	washy	colored
Quantity	broad	-	latter
Appearance	beautiful	grisly	tentacular
Material	waterproof	erose	tabular
Shape	-	-	jagged
Touch	setose	spiny	calorific
Taste	sweet	disgustful	caffeinic
Dimension	stately	wide	graduated
Chronologic	new	-	immutable
Geographic	-	homeless	eastern

Table 3: Example of adjectives' properties

Based on their characteristics have been considered adverbs of manner, adverbs of place, adverbs of time, adverbs of quantity or degree, of affirmation, negation or doubt (grouped as AND adverbs), adverbs as intensifiers or emphasizers and adverbs used in adversative and in consecutives sentences, as listed in Table 2. Only the adverbs of manner may be positive or negative. The adverbs of degree give the idea about the intensity with which something happens or have an impact on sentiment intensity. The others give additional information to the analysis related to the location or the direction or the time. The above tables show more in details the infor-

mation about properties and qualities related to adverbs and adjectives. The above tables show more in details the information about properties and qualities related to adverbs and adjectives. The numbers are referred to pairs of term/synset.

Table 3 describes some examples of the sens- es related to the adjectives properties in order to give an idea of the associations among them. It is possible to notice for instance that the term beau- tiful is associated both to the quality Weather and Appearance, but with different meanings as de- scribed by the WordNet gloss.

3.2 Categorization of Adjectives and Ad- verbs

A semantic categorization task has been per- formed in order to improve the Word Sense Dis- ambiguation of adjectives and adverbs and the meaning concordance with feature terms. The categorization has been made on the gloss asso- ciated to each synset and the result is a set of the main three categories, a subset of WordNet Do- mains, with their related weight. The semantic categorization of the gloss of the synsets permits to relate feature terms and adjectives and/or ad- verbs through a function that defines the seman- tic distance between the feature and the most probable meaning of the adjectives or adverbs related. The same function is used in the WSD of feature terms.

3.3 Feature Extraction

The linguistic resource and the categorization task have been introduced as relevant elements involved in the feature extraction process de- scribed in the following.

The Semantic Classifier categorizes the corpus of reviews and defines, through a resulting set of categories and their weights, the context for the first set of candidate features, extracted by means of a tf-idf function calculated on the sentences having polarity valence. The distinction between sentences having polarity valence from the others has been done using the properties of the lexical resource. The same properties help to better iden- tify the sentiment expressed in relation to the extracted features and are also useful in order to build better meanings relations between the can- didate features and the adjectives and the ad- verbs. The WSD reduces the list of candidate features to a more restricted set having synsets mapped to the top categories of the corpus. The semantic distance function performs the WSD phase in order to identify the correct meanings of feature terms.

The function is a modified version of the Leacock and Chodorow (1998) one, and calculates the semantic distance between the synsets related to the features, defining a matrix of their relations with a weight associated.

The function is part of the algorithm that creates a matrix of relations between features in order to put in evidence and group features in different areas and in more specific thematic set, with the introduction of thresholds' values, or ranges of values, on the calculated weights.

More in details, the function assign a weight to each possible pairs of features, according to the mapping of their synsets to the set of the topic categories. The values are calculated on the number of synset for each feature term. The function, also, provides a value based on the common categories of the synsets and calculates the semantic distance between them by means of an implemented version of the Leacock-Chodorow algorithm.

The matrix has as rows and columns the names of the extracted features and, as values, the weights calculated on each possible couple of features.

3.4 Referring Adjectives and Adverbs to Features

In order to assign an opinion evaluation related to each feature, the collection of sentences expressing a polarity orientation has been used, analyzing the valence expressed by the included adjectives and adverbs. To achieve this step, we used TreeTagger (Schmid, 1994), a syntactic parser able to annotate text with part-of-speech tags and lemma information. TreeTagger also performs a phrase chunking process, identifying into each sentence its sub-constituents such as noun chunks.

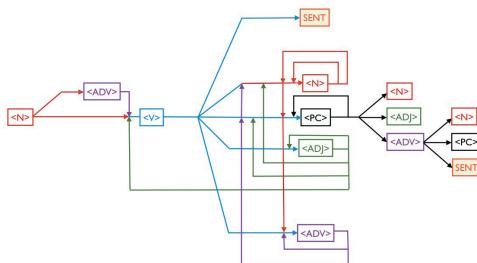


Figure 1. The possible patterns of chunks

The graph has been implemented in order to define all possible sequences of chunks in a set of sentences selected, as showed in the Figure 1.

The graph guided the definition of a Java class that puts in relation the different chunks and refers to the parts of speech in order to have a precise association between the features and their related information.

In the Figure 1 N stands for noun, ADJ for Adjective, ADV for Adverb, V for Verb, PC for prepositional chunk, and SENT is the symbol used to indicate the conclusion of the sentence.

The use of the chunker makes easier the production of a feature-based summary of opinions and a better performance in the definition of the relations between adjective, adverbs and the features related. The semantic distance function has been here again used in order to perform a WSD for the features and the adjectives and adverbs related, by means of their synsets and related categories. The evaluation of the polarity of the meanings of the adjectives and adverbs has been made in order to give a detailed opinion summarization based on each specific feature.

4. Future Works and Conclusion

Several Opinion Mining methods and techniques have been developed in order to analyze contents and reviews. In this paper an extension of WordNet synsets related to adjectives and adverbs, according to a set of properties of relevant importance in the analysis of reviews, has been proposed. The described work is part of a linguistic approach to the automatic extraction of feature terms in Opinion Mining. Moreover, with the introduction of the synsets and the semantic categorization, instead of considering only the words as keywords, we aim to define a method of extraction of more accurate meanings and features from textual resources.

Future works will provide the definition of a tool for the navigation of features and related opinions and the generalization of the approach in order to apply it to general contexts. The tool will perform opinion monitoring activities, an essential task in listening to and taking advantage from consumer preferences and opinions. A validation to support the value of the expressed ideas will be one of the goals of the above mentioned approach and experimental results will be product.

References

- Rodrigo Agerri, Ana García-Serrano, A., 2010. Q-WordNet: Extracting polarity from WordNet senses. Seventh Conference on International Language Resources and Evaluation (LREC 2010).

- Manuela Angioni, Roberto Demontis, Franco Tuveri, 2008. *A Semantic Approach for Resource Cataloguing and Query Resolution*. Communications of SIWN. Special Issue on Distributed Agent-based Retrieval Tools.2010).
- Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Luis Padró, Muntsa Padró, 2006. *FreeLing 1.3: Syntactic and semantic services in an open-source NLP library*. In Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA, Genoa, Italy. <http://nlp.lsi.upc.edu/freeling>.
- Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, 2010, *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*. In Proceedings of LREC-10, 7th Conference on Language Resources and Evaluation, Valletta, MT, 2010, pages 2200-2204
- Farah Benamara, Carmine Cesarano, Antonio Picarillo, Diego Reforgiato, Venkatramana S. Subrahmanian, 2007. *Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone*. In Proceedings of ICWSM 07 International Conference on Weblogs and Social Media, pp. 203-206.
- Xiaowen Ding, Bing Liu, Philip S. Yu, 2008. *A Holistic Lexicon-Based Approach to Opinion Mining*. WSDM '08 Proceedings of the international conference on Web search and web data mining, ACM New York, NY, USA.
- Andrea Esuli, Fabrizio Sebastiani, 2007. *PageRanking WordNet synsets: An application to Opinion Mining*. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics Volume: 45, Issue: June, Publisher: Association for Computational Linguistics, Pages: 424-431.
- Claudia Leacock, Martin Chodorow, 1998. *Combining local context and WordNet similarity for word sense identification*. Fellbaum 1998, pp.265-283.
- Dongjoo Lee, Ok-Ran Jeong, Sang-Goo Lee, 2008. *Opinion Mining of customer feedback data on the web*. In ICUIMC '08 Proceedings of the 2nd international conference on Ubiquitous information management and communication.
- Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, Alfio Gliozzo, A., 2002. *The Role of Domain Information in Word Sense Disambiguation*. Natural Language Engineering, special issue on Word Sense Disambiguation, 8(4), pp. 359-373, Cambridge University Press.
- George A. Miller, 1995. *WordNet: A Lexical Database for English*. Communications of the ACM Vol. 38, No. 11: 39-41.
- Vassiliki Rentoumi, George Giannakopoulos, 2009. *Sentiment analysis of figurative language using a word sense disambiguation approach*. In International Conference on Recent Advances in Natural Language Processing (RANLP 2009), Borovets, Bulgaria, The Association for Computational Linguistics.
- Helmut Schmid, 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In Proceedings of the International Conference on New Methods in Language Processing, pp. 44-49.
- Alessandro Valitutti, Carlo Strapparava, Oliviero Stock, 2004. *Developing affective lexical resources*. Psychnology: 2 (1).
- Piek Vossen (ed), 1998. EuroWordNet: *A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht.

Teaching WordNet to Sing like an Angel and Cry like a Baby: Learning Affective Stereotypical Behaviors from the Web

Tony Veale

Division of Web Science & Technology
Korea Advanced Institute of
Science and Technology (KAIST)
tony.veale@gmail.com

Abstract

Just as words have the potential to mean different things in different contexts, so too can their affective intent vary from one context to another. Thus, in some contexts one might feel complimented to be described as *cunning*, but feel aggrieved and insulted to be so described in another. As concepts become more complex and multifaceted, and accrete more layers of stereotypical associations, their ability to assume different affective profiles in different contexts also increases. Complimentary uses of the stereotype *baby*, for instance, will emphasize the positive behaviors of babies, while insulting uses have many negative behaviors to draw upon and accentuate. In this paper we propose a two-level organization for the affective lexicon, one that can work well with different WordNets. At the first level, stereotype-denoting terms are associated with a rich and nuanced description of their potential behaviors; at the second level, these behaviors are mapped to their affect profiles. This current paper focuses primarily on the first level.

1 Introduction

Context exerts a powerful yet largely unseen influence on our interpretation of natural language utterances. It is context that primes our expectations, to focus our attention on just those senses of a word that are relevant to its linguistic and pragmatic setting. In this way, context successfully hides from us the true complexity of the words we use, and so it can be a surprising experience to open a dictionary, or browse a WordNet (Fellbaum, 1998), and see just how many differ-

ent meanings and nuances a word can convey. The same can be said of a word's *affect*: in context, a word seems to mean just what it is intended to mean, and carry just the right emotional overtones and mood. But viewed out of context, the mapping of words to affect is never quite so direct. Just as words can have many senses, so too can they have a multiplicity of affective uses.

The sense inventories that lexicographers compile for a polysemous word offer a good approximation of the word's potential to convey meaning, but affect can operate across sense boundaries and even within individual senses, at the sub-sense level. Consider the word "baby", used to denote a human infant. In some contexts the word carries a positive affect: babies can be cute and adorable, curious and trusting, and an obvious target of love and affection, especially when asleep. Crying babies, however, can be selfish, whining, drooling, hissing, tantrum-throwing little monsters. Both views are stereotypical of human babies, and either can be intended when a speaker uses the term "baby" figuratively, whether to describe a beloved partner or an annoying colleague. This is a matter of conceptual perspective, not of lexical sense, and many other words exhibit a similar affective duality; "teenager" for instance can mean "whining brat" just as easily as "growing adolescent". The concepts *Baby* and *Teenager* are complex and multifaceted, and different uses in context may highlight different stereotypical behaviors of each. Their affective meaning in context is therefore not so much a function of which lexical sense is intended but of which behaviors are highlighted, and of the perceived affect of those behaviors.

Before we can build a nuanced model of affect for a lexical resource like WordNet, we first need to understand the stereotypical behaviors on which affect is determined. With a sufficiently

rich behavioral model, we can determine the affect of a word like “*baby*” or “*teenager*” on a case-by-case and context-by-context basis, rather than wiring a one-size-fits-all measure of average affect directly into the lexicon. In short, we propose a two-level structure for a context-sensitive affective lexicon: a mapping of word-concepts to their normative stereotypical behaviors (e.g. *mewling*, *shrieking*, *drooling*, *sleeping* and *smiling*); and an affective profile of those behaviors (e.g. indicating the degree to which *shrieking* is unpleasant and *smiling* is pleasant). The affect of a word/concept in context can then be calculated as a function of the affect of its stereotypical behaviors that are primed in that context.

In this paper we focus on the first stage of the model – the construction of a rich behavior-net that associates stereotypical concepts with their expected behaviors. This stage will serve as the foundation for a subsequent model of context-sensitive lexical affect. We start in section 2 with a survey of related work in the area of stereotypes and affect, before outlining our current approach in section 3. We report on the scale of this work, and its current state, in section 4, and conclude the paper in section 5 with a brief preview of the next stage of construction for our behavior-based model of lexical affect.

2 Related Work

In its simplest form, an affect lexicon assigns an affective score – along one or more dimensions – to each word or sense. Whissell’s (1989) *Dictionary of Affect*, for instance, assigns a trio of numeric scores to each of its 8000+ words to describe three psycholinguistic dimensions: *pleasantness*, *activation* and *imagery*. In the DoA, the lowest pleasantness score of 1.0 is assigned to words like *abnormal* and *ugly*, while the highest, 3.0, is assigned to words like *wedding* and *wining*. Less extreme words are assigned pleasantness scores closer to the DoA mean of 1.84. Whissell’s DoA is based on human ratings, but Turney (2002) shows how such scores can be assigned automatically, using statistical measures of word association in web text.

For reliable results on a large-scale, Mohammad & Turney (2010) used the Mechanical Turk to elicit human ratings of the emotional content of different words. Ratings were sought along the eight primary emotional dimensions identified by Plutchik (1980): *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust*. Automated tests were used to exclude unsuitable raters, and

in all, 24,000+ word-sense pairs were annotated by five different raters. Thus, words that suggest fearful contexts, like *threat*, *hunter* and *acrobat*, are all assigned a significant score on the *fear* dimension, while *disease* and *rat* score highly on the *disgust* dimension.

Strapparava & Valitutti (2004) provide a set of affective annotations for a subset of WordNet’s synsets in a resource called *Wordnet-affect*. The annotation labels, called *a-labels*, focus on the cognitive dynamics of emotion, allowing one to distinguish e.g. between words that denote an *emotion-eliciting situation* and those than denote an *emotional response*. Esuli & Sebastiani (2006) also build directly on WordNet as their lexical platform, using a semi-supervised learning algorithm to assign a trio of numbers – *positivity*, *negativity* and *neutrality* – to word senses in their newly derived resource, *SentiWordNet*. (Note that *Wordnet-affect* also supports these three dimensions as a-labels, as well as a fourth, *ambiguous*). Esuli & Sebastiani (2007) improve on their affect scores by running a variant of the PageRank algorithm (see Mihalcea and Tarau, 2004) on the implicit graph structure that tacitly connects word-senses in WordNet via their textual glosses.

These lexicons attempt to capture the affective profile of a word/sense when it is used in its most normative and stereotypical guise, but they do so without an explicit model of stereotypical meaning. Veale & Hao (2007) describe a web-based approach to acquiring such a model. They note that since the simile pattern “as ADJ as DET NOUN” presupposes that NOUN is an exemplar of ADJness, it follows that ADJ must be a highly salient property of NOUN. Veale and Hao thus harvested tens of thousands of instances of this pattern from the web, to extract sets of properties (ADJ) for thousands of commonplace terms (NOUN). They show if one estimates the pleasantness of a term like *snake* or *artist* as a weighted average of the pleasantness of its properties (like *sneaky* or *creative*) in a resource like Whissell’s DoA, the estimated scores show a reliable correlation with the DoA’s own scores. In other words, it makes computational sense to calculate the affect of a word-concept as a function of the affect of its most salient semantic properties.

These differing approaches are reconciled in the two-level model outlined here. A variant of the approach in Veale & Hao (2007) is used to acquire a model of stereotypical behaviors from the web. The affective profile of these behaviors

can be described using any of the above approaches, such as DoA or SentiWordNet. Only behaviors (level 2) are pre-assigned affective scores in the lexicon; for entities exhibiting these behaviors (level 1), affect is calculated on demand and in context.

3 Learning Behaviors from the Web

Veale & Hao make the simplifying but unjustified assumption that all stereotypical properties are adjectival in nature, and work from adjectival properties (as inventoried by WordNet) to the nouns that exemplify them by successively binding *ADJ* in the web query “as *ADJ* as a *NOUN*” to different adjectives. The resulting enfilade of queries is sent in rapid succession to the search engine Google. All bindings for *NOUN* are then automatically extracted from the results before being manually inspected. Here we instead use the *like* patterns “*VERB+ing like a NOUN*” and “*VERB+ed like a NOUN*”, the preferred simile patterns to describe behavior.

Before performing a large-scale trawl of the web, we first conduct a pilot study on the Google n-grams (Brants & Franz, 2006), a database of contiguous n-word strings ($1 \leq n \geq 5$) with a web frequency of 40 or higher. The pattern “*VERB+ing like a NOUN*” matches over 8,000 4-grams, while “*VERB+ed like a NOUN*” matches almost 4,000. However, we find here a good deal of empty behaviors, such as *acting* (as in “*acting like a baby*” rather than “*acting like an actor*”) and *looking* (as in “*looking like a fool*”). Indeed, just three empty behaviors – *acting*, *looking/looked* and *seemed* – account for almost 2,000 n-gram matches. Others, like *walking* and *eating*, are too general and merely allude to a stereotypical behavior (as in “*walking like a penguin*”) rather than explicitly providing the specific behavior (*waddling*). Panning the n-gram matches yields a few hundred nuggets of stereotypical insight, such as “*circling like a shark*”, “*salivating like a dog*” and “*clinging like a leech*”. Our pilot study reveals that most instances of the *like*-simile patterns are not so specific and informative, making a large-scale web trawl with these patterns impracticable.

Instead we use a hypothesis-driven approach by first looking for attested mentions of a specific behavior with a given noun. Consider the target noun *zombie*: searching the Google 3-grams for matches to the patterns “*DET VERB+ing zombie*” and “*DET VERB+ed zombie*” yields the following hypotheses for the stereotypical behav-

ior of zombies (numbers in parentheses are the frequencies of the corresponding 3-grams):

```
{ decomposing(1454), devastating(134), shambling(115),
rotting(103), ravaged(98), brainwashed(94),
drooling(84), freaking(83), attacking(80), crazed(79),
obsessed(73), infected(72), marauding(71),
disturbed(65), wandering(64), reanimated(54),
flying(52), flaming(52), revived(47), decaying(41),
unexpected(40)}
```

For each attested behavior we generate the corresponding *like*-simile, such as “*decomposing like a zombie*”, and determine its frequency on the open web. Corresponding non-zero frequencies, obtained using Google, are as follows:

```
{ drooling(4480), wandering(3660), shambling(1240),
revived(860), rotting(682), brainwashed(146),
reanimated(141), infected(72), flaming(52),
decaying(46), decomposing(8), attacking(7), flying(6),
freaking(2), obsessed(3)}
```

Unlike Veale & Hao then, we do not use a relatively small (~ 2000) set of queries that are made wide-ranging through the use of wild-cards, but generate a very large set of specific queries (with no wild-cards) that each derive from an attested combination of a specific behavior and a specific noun. We are careful not to dispatch queries that contain empty behaviors, a list of which is determined during our initial pilot study with the Google n-grams. In all, we dispatch over 500,000 web queries, for the same number of attested combinations. No parsing of the web results is needed, and we need record only the total number of returned hits per query / combination.

4 Initial Evaluation

The 3-gram patterns “*DET VERB+ing NOUN*” and “*DET VERB+ed NOUN*” attest to the plausibility of a given noun-entity exhibiting a specific behavior, but they are only weakly suggestive about what is actually typical. As a basis for generating hypotheses about stereotypical behavior these patterns over-generate significantly, and less than 20% of our queries yield non-zero result sets when sent to the web.

As shown by the *zombie* example above, some web-attested behaviors are best judged as idiosyncratic rather than stereotypical. While *rotting*, *decaying* and *shambling* are just the kind of behaviors we expect of zombies, *freaking*, *flying* and *flaming* are ill-considered oddities that our behavior model can well do without. As one

might expect, such oddities tend to have lower web frequencies than more widely-accepted behaviors (like *drooling*), yet as noted in Kilgarriff (2007), raw web frequencies can be an unreliable guide to what is typical. Note for instance how *decomposing* has a low frequency of just 8 uses on the web (as indexed by Google).

Our web data exhibits another interesting phenomenon. Consider the noun-entities for which the behavior *brainwashed* is attested, both in the 3-grams (“*a brainwashed NOUN*”) and on the web (“*brainwashed like a NOUN*”):

```
{ cult(1090), zombie(146), robot(9), child(7), fool(4),
  kid(4), idiot(3), soldier(2) }
```

Since cults often use brainwashing, we can consider *cult* to be stereotypical for this behavior. Zombies and robots, however, are not typically brainwashed, nor indeed are they even brainwashable. Rather, it is more accurate to suggest that the victims of brainwashing often resemble *robots* and *zombies*, and to the extent that brainwashing is made possible by being weak-minded, *fools*, *idiots*, *kids* and *children*. This appears to be an example of what Bolinger (1988) dubs *ataxis*, insofar that *brainwashed* is a “migrant modifier” that more aptly describes the target of the simile than the vehicle (*robot* or *zombie*). In this case we can sensibly conclude that *brainwashed* is a figurative behavior of *robots* and *zombies* (since they typically act like a brainwashed person) and is the kind of association we want in our behavioral model. In contrast, it would not be sensible to include *brainwashing* as part of the behavioral description of *fools*, *idiots*, *kids*, *children* or even *soldiers* (though the latter is perhaps debatable).

Ultimately, the stereotypicality of a behavioral association is a pragmatic gut issue for the designer of a lexico-semantic resource, one that cannot be automatically resolved by considering web frequency (or other statistical quantities) alone. As with the design of WordNet itself, it is best resolved by asking and answering the question “is this an association that I would want in my lexicon?”. For this reason, we filter the results of the web harvesting process manually, to ensure that the final model contains only those behavioral descriptions that a human would consider typical. In the end then, our approach is a semi-automatic one: automated processes scour the Google n-grams for behavioral hypotheses, and seek supporting evidence for these hypotheses on the web (in the form of *like*-similes), be-

fore a manual pass is finally conducted to ensure the model has the hand-crafted quality of a resource like WordNet.

This semi-automation allows us to build a behavioral model of high quality and significant scale. The model maps 5649 unique nouns to 4256 unique behaviors and contains approx. unique 44,000 mappings overall. This behavior-based model is thus more than three times larger than the adjectival stereotype model reported in Veale & Hao (2007), which contains just over 12,000 noun-to-adjective mappings.

5 Next Steps

The behavioral model, which captures the stereotypical behavior of thousands of word-concepts from *apes* to *zombies*, can be viewed as a complementary addition not just to WordNet but to the other knowledge resources previously described. Most obviously it complements the adjectival-stereotype model of Veale & Hao, and integrating the two would yield a larger and richer resource, of stereotypical descriptions that combine both adjectival and behavioral properties. For example, in a combined model, the *baby* stereotype has the following 163 properties:

```
{ delicate, squalling, weeping, baptized, adopted, startled,
  attentive, blessed, teeny, rocked, adorable, whining,
  bundled, toothless, placid, expected, rescued, treasured,
  new, sleepy, indulged, slumbering, weaned, supple,
  helpless, small, sleeping, animated, vulnerable, wailing,
  cradled, kicking, soft, rested, bellowing, blameless,
  grinning, screaming, tiny, cherished, reliant, thriving,
  loveable, guileless, mute, inexperienced, dribbling,
  unthreatening, nursed, angelic, bawling, beaming, naked,
  spoiled, scared, weak, squirming, blubbering, contented,
  smiling, wiggling, mewling, blubbing, sniffing, overtired,
  dimpled, loving, dear, tired, powerless, bewildered,
  peaceful, distressed, naive, wee, soiled, sucking, fussy,
  gurgling, vaccinated, heartwarming, pouting, constipated,
  drooling, quiet, wiggly, lovable, bare, weaning, suckling,
  cute, bald, whimpering, tender, pampered, incontinent,
  fleshy, charming, dependent, artless, fussing, flabby,
  babbling, warm, giddy, crawling, snoozing, hairless,
  cuddled, sweet, sobbing, squealing, wrapped, cooing,
  swaddled, laughing, toddling, fragile, innocent, moaning,
  gentle, terrified, precious, cranky, giggling, confused,
  cuddly, fat, ignorant, snoring, young, howling, screeching,
  shrieking, trusting, shivering, napping, resting,
  frightened, fresh, loved, demanding, chubby, adored,
  appealing, happy, tame, relaxed, wriggly, rocking,
  wriggling, conceived, clean, content, smooth, crying,
  submissive, bumbling, pink, sniveling, orphaned,
  harmless, pure }
```

A cursory glance at this list reveals a rich description of the stereotypical baby, one that incorporates pleasant and unpleasant behaviors in ample numbers. It makes little sense to reduce such a nuanced description to a single measure of gross lexical affect, or to parcel the description into separate senses, each with its own subset of behaviors. Instead, the partitioning of the description can be done on demand, and in context, to suit the speaker's meaning: if a term is used pejoratively, we focus on those qualities that are typically unpleasant (*sniveling*, *submissive*, *cranky*, *whimpering*, etc.); if the term is used affectionately, we focus instead on those that typically convey affection (*blessed*, *delicate*, *pure*, *loved*, *trusting*, etc.); and so on. The affective rating of different qualities can be ascertained from any of the existing resources discussed earlier, with more or less success. Whissell's DoA is perhaps the most limited, while Mohammad & Turney's eight-dimensional model of emotion seems to possess the most nuance and power.

However, even basic properties and behaviors can be construed differently from one context to another. In some settings, for instance, *cunning* may be a positive description; in most others, it will likely be seen as negative. Many adjectival properties exhibit this duality of affect, such as *proud*, *tough*, *tame* and *fragile*, and the description of the stereotypical baby above contains many that could be used to compliment in one context and to insult in another.

For this reason, we shall concentrate next on the construction of a nuanced model of behavioral affect, in which the affective profile of a behavior or adjectival property (and thus of the entity that exhibits that behavior in context) changes in response to the intended meaning of the speaker. This model, which will form the second stage of the two-level affective lexicon outlined in the introduction, will allow us to see the positive in properties like *trusting*, *cunning* and *demanding*, and the negative in properties like *proud*, *unthreatening* and *innocent*, as the context demands.

The behavior model described here will be of considerable use in this goal, since we now have a reliable, large-scale means of determining which properties and behaviors co-occur with which. For instance, the baby stereotype tells us that *sniveling* co-occurs with *submissive* and *cranky* co-occurs with *whimpering*. From these co-occurrence patterns we have constructed a weighted graph of mutually-supporting behaviors and the entities that exhibit them. We are now

conducting experiments on the use of PageRank and other graph-theoretic algorithms (as used in Rada & Tarau, 2004; Esuli & Sebastiani, 2007) to identify the most effective means of exploiting graph structure in the determination of affect.

6 Conclusions (for now)

The availability of large-scale lexical resources with rich sense inventories, like WordNet, has made it possible to move from affect lexica that assign gross affective properties directly to words (e.g., Whissell's DoA) to more sophisticated organizations that assign affect to particular word senses only (e.g., Wordnet-affect and SentiWordNet). This allows an affect lexicon to tease apart the aspects of a word/concept that carry positive or negative connotations (such as the indiscriminate and clumsy senses of *butcher*, or the heroic sense of *hero*, but not the sandwich sense of *hero*) and carefully assign the right properties to just the right senses.

But affect is not a phenomenon that respects sense boundaries, and the affective connotation of one sense of a word can easily spread to others. Thus, all senses of the word *butcher*, including the purely literal sense of a professional meat vendor, are tainted by the negative connotation of the metaphoric extension that describes an indiscriminate murderer. Likewise, the same sense of a word can be used with different affective connotations in different contexts, because even individual senses – what Cruse (1986:49) conceives of as “unitary ‘quanta’ of meaning” – denote complex objects with their own wide ranges of typical properties and expected behaviors. While sense distinctions allow us to make our affect lexica more precise, sense boundaries do not demarcate affect boundaries as surely as we would like. But the solution does not lie in sense proliferation, in which even more fine-grained senses are added to WordNet and other resources. Rather, it lies in an ability to dynamically construe new perspectives on existing senses as the context demands.

The work reported here is just one step in this direction. Only by adequately modeling what is typical and salient – that is, what is *stereotypical* – of the entities denoted by our words and their senses, can we begin to model how speakers in context subtly shift the boundaries of sense to effectively communicate an affective meaning.

7 Acknowledgements

This research was supported by the WCU (World Class University) program under the National Research Foundation of Korea, and funded by the Ministry of Education, Science and Technology of Korea (Project No: R31-30007).

References

- Dwight Bolinger. (1988). Ataxis. In Rokko Linguistic Society (ed.), *Gendai no Gengo Kenkyu (Linguistics Today)*, Tokyo:1—17.
- Thorsten Brants and Alex Franz. (2006). *Web IT 5-gram Version 1*. Linguistic Data Consortium.
- D. A. Cruse. (1986). *Lexical Semantics*. London: Cambridge University Press.
- Andrea Esuli and Fabrizio Sebastiani. (2006). Senti-WordNet: A publicly available lexical resource for opinion mining. *Proceedings of LREC-2006, the 5th Conference on Language Resources and Evaluation*, 417-422.
- Andrea Esuli and Fabrizio Sebastiani. (2007). PageRanking WordNet Synsets: An application to opinion mining. *Proceedings of ACL-2007, the 45th Annual Meeting of the Association for Computational Linguistics*.
- Christiane Fellbaum (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.
- Adam Kilgarriff. (2007). Googleology is Bad Science. *Computational Linguistics*, 33(1):147-151.
- Rada Mihalcea and Paul Tarau. (2004). TextRank: Bringing Order to Texts. *Proceedings of EMNLP-04, the 2004 Conference on Empirical Methods in Natural Language Processing*.
- Saif M Mohammad and Peter D. Turney. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotional lexicon. *Proceedings of the NAACL-HLT 2010 workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles, California.
- Robert Plutchik. (1980). A general psycho-evolutionary theory of emotion. *Emotion: Theory, research and experience*, 2(1-2):1-135.
- Peter D. Turney. (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". *Proceedings of ACL-2002, the 40th Annual Meeting of the Association for Computational Linguistics*. pp. 417-424.
- Carlo Strapparava and Alessandro Valitutti. (2004). Wordnet-affect: an affective extension of Wordnet. *Proceedings of LREC-2004, the 4th International Conference on Language Resources and Evaluation*, Lisbon.
- Tony Veale and Yanfen Hao. (2007). Making Lexical Ontologies Functional and Context-Sensitive. *Proceedings of ACL-2007, the 45th Annual Meeting of the Association of Computational Linguistics*, 57-64.
- Cynthia Whissell. (1989). The dictionary of affect in language. In R. Plutchik and H. Kellerman (Eds.) *Emotion: Theory and research*. Harcourt Brace, 113-131.

Multiword Verbs in WordNets

Veronika Vincze¹, Attila Almási² and János Csirik¹

¹Hungarian Academy of Sciences, Research Group on Artificial Intelligence

Tisza Lajos krt. 103., 6720 Szeged, Hungary

{vinczev, csirik}@inf.u-szeged.hu

²University of Szeged, Department of Informatics

Árpád tér 2., 6720 Szeged, Hungary

vizipal@gmail.com

Abstract

In this paper, we describe how wordnets treat multiword verbs. We pay special attention to the English and Hungarian wordnets and we argue that from a multilingual perspective it is recommended to store idioms and light verb constructions as a whole rather than listing their parts separately. In order to enhance their applicability in multilingual applications, a unified treatment should be applied for subtypes of multiword verbs.

1 Introduction

In natural language processing, one of the most challenging tasks is the proper treatment of multiword expressions (MWEs). Multiword expressions are lexical items that can be decomposed into single words and display lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy (Sag et al., 2002; Kim, 2008; Calzolari et al., 2002). To put it differently, they are lexical items that contain space or “idiosyncratic interpretations that cross word boundaries”. Multiword expressions are frequent in language use and they usually exhibit unique and idiosyncratic behavior, thus, they often pose a problem to NLP systems.

In this paper, we describe how wordnets treat multiword expressions and we pay special attention to multiword verbs. Multiword verbs comprise phrasal verbs, light verb constructions and idioms¹, how-

¹Idioms usually consist of a verb phrase and they are semantic predicates, thus, their grammatical function is similarly to that of verbs. This is why we consider them as a subtype of multiword verbs.

ever, we focus on idioms and light verb constructions in our investigations as they represent two different levels of compositionality: while idioms are totally non-compositional, light verb constructions are semi-compositional (i.e. the meaning of the noun plays an important role in computing the meaning of the whole structure). We concentrate on English and Hungarian and we argue that from a multilingual perspective, it is more advisable to store multiword expressions as a whole rather than listing their parts separately.

The structure of the paper is as follows. First, the decomposability of multiword expressions is discussed, then it is shown how idioms and light verb constructions should be treated in wordnets. The paper concludes with a comparison of methods offered for these two types of multiword verbs.

2 The decomposability of multiword expressions

Multiword expressions can be classified according to their semantic decomposability (Sag et al., 2002; Nunberg et al., 1994). If the parts of the MWE can be interpreted as having a special sense unique to this construction, that is, there can be a word-to-word mapping between the lexical and the semantic level, it is called a decomposable MWE. An English and a Hungarian example are offered here:

to spill the beans

‘to reveal a secret’

beans = ‘secret’

spill = ‘reveal’

veszi a lapot
take-3SGOBJ the card-ACC
'to understand the message'

vesz = 'understand'
lap = 'message'

It should be noted that in the English example, the definite article in the idiom corresponds to an indefinite one on the semantic level, however, all words in the idiom can be mapped to another one on the semantic level. If no such correspondence can be found, the MWE is considered to be non-decomposable. An example is *to bite the dust* 'to die' or its Hungarian equivalent *fűbe harap* (grass-ILL bites) which meaning cannot be decomposed in a way to match the single words within the expression.

The above distinction may have interesting implications for wordnet building. If the parts of a MWE can be attributed a special distinct meaning, the question arises whether this meaning should be added to the sense inventory of the given words or not, in other words, to decompose its meaning or not. From another perspective, should decomposable MWEs be stored as one unit in wordnets (i.e. as one synset) or should their parts be separately listed in synsets with the corresponding senses? In order to answer this question from a multilingual aspect, we first examine how the Princeton WordNet (PWN) (Miller et al., 1990) and the Hungarian WordNet (Miháltz et al., 2008) treat multiword verbs.

3 Idioms in the English and the Hungarian wordnets

We can find the following synset in PWN:

{gutter:2, sewer:3, toilet:3}

These literals are parts of idioms, which are not listed as a whole in PWN. The PWN synset means "misfortune resulting in lost effort or money", however, it is not obvious from the representation that this sense is valid only within the idiom, i.e. in combination with *go* or *be* and a preposition.

The Hungarian equivalent of the above synset is a non-lexicalized one²:

²Creating the HuWN database practically meant rendering the PWN synsets into Hungarian, that is, Hungarian equivalents

(WC, ablak, csatorna; kidobhatod az ablakon) 'toilet, window, gutter; you can throw it out the window'

Thus, it seems that the above PWN synset has no lexicalized Hungarian counterpart although there are Hungarian idioms that express the same meaning, e.g. *kidobhatja az ablakon* (out.throw-MOD-3SGOBJ the window-SUP) or *lehúzhatja a WC-n* (down.flush-MOD-3SGOBJ the toilet-SUP). Thus, it would have been feasible to create a Hungarian synset with the nominal parts of the idioms such as:

(WC, ablak, csatorna) 'toilet, window, gutter'

As Osherson and Fellbaum (2010) propose, the connection between the parts of idioms can be signaled by idiom-specific relations between synsets. However, the major problem with this approach is that not all members of the synset can be paired with the same verb: for instance, in Hungarian, there are no phrases like **lehúzhatja az ablakot* 'to flush the window' or **kidobhatja a WC-n* 'to throw it out the toilet'. Thus, it would be complicated to signal which literal can be paired with which verb if the nominal parts of the idioms with similar meanings are included in the very same synset.

From a multilingual perspective, it is interesting to note that most multiword expressions have an equivalent in other languages, however, it may well be the case that the linguistic structure of the MWE with the same meaning in two languages do not coincide or one of them is decomposable (the parts of the MWE can be interpreted as having a special sense unique to this construction, that is, there can be a word-to-word mapping between the lexical and the semantic level) and the other one is not as in:

to be on cloud number 9

*örül, mint majom a
be.glad as monkey the
farkának
tail-3SGPOSS-DAT*
'to be extremely happy'

had to be found for PWN synsets. Whenever this was not possible, e.g. due to differences in culture, language use or grammar, the synset was marked as *non_lexicalized* in Hungarian and an approximate definition was given for the English concept.

Here the English idiom is decomposable – *cloud number 9* corresponds to *happiness* – while in Hungarian, the verb *örül* ‘be glad’ corresponds to the “happy” component in the meaning of the idiom, however, *majom* ‘monkey’ and *farkának* ‘to his tail’ cannot be matched to any meaning component. On the other hand, in English, *cloud nine* is listed in a synset denoting happiness (bliss:1, blissfulness:1, cloud nine:1, seventh heaven:1, walking on air:1) in accordance with the proposal found in Osherson and Fellbaum (2010) but in Hungarian, no mention of the idiomatic usages is made in the corresponding synset (*elragadtatás:2, mennyei boldogság:1, üdvösségek:1*), although *hetedik mennyei boldogság* ‘seventh heaven’ could have been listed here as there is an idiom with a similar meaning (*a hetedik mennyei boldogságban érzi magát* (the seventh heaven-INE feel-3SGOBJ himself-ACC) ‘to be in seventh heaven’). However, none of the components of the idiom *örül, mint majom a farkának* could be included in this synset in HuWN since there is no noun in the idiom corresponding to *cloud nine* that can be included in the nominal hierarchy. This also highlights that the treatment of idioms is somewhat problematic in HuWN: sometimes, synsets corresponding to idiom parts in PWN are marked as *non-lexicalized* in HuWN, or no idiom parts are mentioned within the synset. In order to solve this problem, we propose to include the whole idiom as a lexical unit in the verbal parts of wordnets, which can be easily matched to another idiomatic synset in the other language, without being forced to find a nominal component in both languages that have the same meaning within the MWE. Thus, the following synsets can be proposed:

{be in the gutter, go down the sewer, be in the toilet}

{lehúzhatja a WC-n (down.flush-MOD-3SGOBJ the toilet-SUP), kidobhatja az ablakon (out.throw-MOD-3SGOBJ the window-SUP)}

Although Osherson and Fellbaum (2010) suggest that parts of decomposable idioms should be consequently included in wordnets on the basis of the fact that there are idioms with the same or similar meanings, thus, their components may form a single synset (compare *seventh heaven* and *cloud nine*),

they also admit that prepositions and other function word elements of idioms cannot be given in this way since PWN only includes nouns, verbs, adjectives and adverbs. In Hungarian, the situation is somewhat more complicated since nouns get suffixes in sentences, which cannot be signaled in any way by listing only parts (lemmas) of idioms. We argue, however, that including the whole idiom as one lexical unit is more beneficial from the aspect of multilinguality for it is easier to find the other language equivalent of idioms than the equivalent of idiom parts, and, on the other hand, the whole idiom is listed and not only its nominal, adjectival and verbal parts. Nouns in idioms also occur in the right grammatical form (i.e. with the correct suffix). In this way, non-lexicalized synsets related to idiom parts can also be eliminated. On the other hand, decomposable and non-decomposable idioms are treated in the same way: they are both listed as a whole. With this solution, idioms that share the same meaning should be treated similarly to single synonymous words, that is, they can be included within one synset.

4 Multiword verbs in the Hungarian WordNet

In the following, it is shown how multiword verbs are included in the conceptual hierarchy of the Hungarian Wordnet.

Among the 3607 verbal synsets of the Hungarian WordNet, 84 contain at least one multiword verb (106 altogether). Among them, 10 phrases consist of an adjective in the translative case and the verb *tesz* ‘make’, e.g. *jobbá tesz* (better-TRANSL makes) ‘to ameliorate’. The English equivalents of these synsets are typically single verbs one meaning component of which is ‘make’ as it is shown in their definition, for instance:

ID: ENG20-00498510-v
 Synonyms: {disable:1, disenable:1, incapacitate:1}
 Definition: make unable to perform a certain action

In Hungarian, the meaning component ‘make’ is explicitly expressed by the verb *tesz* ‘make’.

Although there are some idiomatic expressions such as *dűlőre jut* (brink-SUB gets) ‘to come to an

agreement' among multiword verbs in HuWn, most of them belong to the category of light verb constructions. Light verb constructions consist of a nominal and a verbal component where the noun is usually taken in one of its literal senses but the verb usually loses its original sense to some extent. The Hungarian WordNet treats them as separate lexical units, that is, they behave as normal literals.

When constructing the Hungarian Wordnet, wordnet builders were given special instructions to include the most frequent light verb constructions in synsets (frequency data were estimated on the basis of the Hungarian National Corpus (Váradi, 2002)). They can be found in synsets together with their verbal counterparts as in:³

ID: ENG20-00777368-v
 Synonyms: {engedélyez:1, **engedélyt ad:1**}
 Definition: Hatóság vagy hivatal engedélyt megad.

ID: ENG20-00777368-v
 Synonyms: {authorize:1, authorise:2, pass:24, clear:4}
 Definition: Grant authorization or clearance for.

Sometimes, there are more than one light verb constructions within one synset, which entails that they are synonyms:

ID: ENG20-00862885-v
 Synonyms: {hálát ad:1, köszönetet mond:1, köszönetet nyilvánít:1, megköszön:1, köszön:1}
 Definition: Köszönetét fejezi ki valakinek valamiért.

ID: ENG20-00862885-v
 Synonyms: {thank:1, **give thanks:1**}
 Definition: Express gratitude or show appreciation to.

³The corresponding English synsets are imported from the Princeton WordNet, thus, they do not always contain a light verb construction and translation is not always word-by-word.

In certain cases, the synset contains only one light verb construction, that is, it is regarded as a separate lexical unit (having one entry in the dictionary or rather forming one synset) in the wordnet):⁴

ID: ENG20-00992244-v
 Synonyms: {szóba hoz:1}
 Definition: Szól róla, megemlíti.

ID: ENG20-00992244-v
 Synonyms: {raise:19, bring up:6}
 Definition: Put forward for consideration or discussion.

Based on these examples, HuWN can be considered as a database in which light verb constructions are treated as separate lexical entries.⁵ However, as wordnets contain several lexical relations among synsets, it would prove useful to link the synset of the nominal component to that of the light verb construction, e.g. the relation *derivative*⁶ might connect them to each other, thus signaling their morphological and semantic interrelatedness (e.g. *engedély* 'permission' is paired with {engedélyez:1, engedélyt ad:1} 'authorize'). This extension of relations between synsets would be fruitful in the sense that the synsets of the construction and its components would be directly connected hence they could inherently be matched without any further steps. From a multilingual perspective, although *make a decision* and *döntést hoz* are translational equivalents, this cannot be deduced without analyzing the definition. In order to enrich the applicability of wordnets in the automatic translation of multiword expressions, we also suggest that light verb constructions be included in wordnets in a more systematic way, i.e. they should be literals within the synset and not only parts of the definition.

⁴However, the definition itself contains single word equivalents of the concept.

⁵As for PWN, light verb constructions are sometimes treated as lexical units (e.g. *give thanks*) but in other cases, it is the definition that contains the light verb construction equivalent of the literals (e.g. {decide:1, make up one's mind:1, determine:5} is defined as "reach, make, or come to a decision about something").

⁶In HuWN, no derivative relations have been included so far.

5 Discussion

We argue that multiword verbs such as idioms and light verb constructions should be listed as one unit in wordnets. In this way, there is no difference in treating decomposable and non-decomposable idioms and other language equivalents of the expressions are easier to find. From a theoretical point of view, this means that multiword expressions are regarded as a separate lexical unit, reflecting the semantic unity of the construction. This is in line with construction grammars (see e.g. Goldberg (1995)), where contents and forms are paired to form a construction with typically unpredictable meaning.

However, there is a difference between idioms and light verb constructions as regards their linking to the synsets of their members. Since the nominal component of light verb constructions preserves its original meaning to some degree, we proposed to connect the nominal component to the synset that contains the light verb construction. On the other hand, in the case of idioms no detectable connection between meanings of the parts of the idiom and the whole phrase can be established that is why we suggest not connecting them.

6 Conclusions

In this paper we have suggested that it is advisable to treat multiword verbs such as idioms and light verb constructions as one unit in wordnets. First, their semantic unity is reflected in this way, second, it is easier to match other language equivalents of the same units. We have also argued that a unified treatment should be applied for types of multiword verbs in order to enhance their applicability in multilingual applications. The revision of idioms and light verb constructions in wordnets is, however, left for future work.

Acknowledgments

This work was supported by the Project “TÁMOP-4.2.1/B-09/1/KONV-2010-0005 – Creating the Center of Excellence at the University of Szeged”, supported by the European Union and co-financed by the European Regional Development Fund.

References

- Nicoletta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1934–1940, Las Palmas.
- Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.
- Su Nam Kim. 2008. *Statistical Modeling of Multiword Expressions*. Ph.D. thesis, University of Melbourne, Melbourne.
- Márton Mihálcz, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi. 2008. Methods and Results of the Hungarian WordNet Project. In Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Fourth Global WordNet Conference (GWC 2008)*, pages 311–320, Szeged. University of Szeged.
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.
- Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70:491–538.
- Anne Osherson and Christiane Fellbaum. 2010. The Representation Of Idioms In WordNet. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors, *Principles, Construction and Application of Multilingual Wordnets. Proceedings of the Fifth Global WordNet Conference (GWC 2010)*, Mumbai, India. Narosa Publishing House.
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.
- Tamás Váradi. 2002. The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 385–389, Las Palmas de Gran Canaria. European Language Resources Association.

Cross-lingual event-mining using wordnet as a shared knowledge interface

Piek Vossen

VU University Amsterdam
p.t.j.m.vossen@vu.nl

Aitor Soroa, Beñat Zapirain, German Rigau

University of the Basque Country
a.soroa@ehu.es
benat.zapirain@ehu.es
german.rigau@ehu.es

Abstract

We describe a concept-based event-mining system that maximizes information extracted from text and is not restricted to predefined knowledge templates. Such a system needs to handle a wide range of expressions while being able to extract precise semantic relations. The system uses simple patterns of linguistic and ontological constraints that are applied to a uniform representation of the text. It uses a generic ontology based on DOLCE and wordnets in different languages to extract events from text in these languages in an interoperable way. The system performs unsupervised domain-independent event-mining with promising results. Error-analysis showed that the semantic model and the mapping of text to concepts through word-sense-disambiguation (WSD) are not the main cause of the errors but the complexity of the grammatical structures and the quality of parsing. Using the same semantic model and their cross-wordnet links, our English event-mining patterns were transferred to Dutch in less than a day's work. The platform was tested on the environment domain but can be applied to any other domain.

1 Introduction

Traditionally, Information Extraction (IE) is the task of filling template information from previously unseen text which belongs to a predefined domain (Peshkin and Pfeffer, 2003). Standard IE systems are based on language-specific pattern matching (Kaiser and Miksch, 2005), consisting of language specific regular expressions and associated mappings from syntactic to logical forms.

The major disadvantage of traditional IE systems is that they focus on satisfying precise, narrow, pre-specified requests e.g. all names of directors of movies. Compared to full text indexes in Information Retrieval (IR), IE systems only cover a small portion of the knowledge in texts while capturing deeper semantic relations.

Alternatively, the KYOTO system¹ combines the comprehensiveness of IR systems with the depths of IE systems. Furthermore, the system can be applied to different languages and domains in an uniform way. It tries to extract any event and its participants from the text, and relate them to time and place. To achieve this, it uses a full text representation format and a wide range of knowledge in the form of wordnets and a generic ontology (Vossen and Rigau, 2010). Word-sense-disambiguation (WSD) is a crucial step to map text to concepts. We implemented a two-phased WSD strategy and show the effects on event-extraction. We first apply a state-of-the-art WSD to words in context, scoring all possible synsets in a wordnet. Each of these synsets is mapped to a shared ontology. From this mapping, all possible ontological implications are derived. Next, the mining system extracts all possible interpretations of all sequences of ontological concepts that match the patterns. In the second-phase, we select an interpretation only if there is a choice using the WSD score. The system has been tested on texts from the environment domain. However, the knowledge resources and patterns are generic and can be applied to any other domain.

In the next section, we describe the general architecture of the KYOTO system and the knowledge structure. In section 3, we describe the module for mining knowledge from the text. In section 4, we describe the evaluation results and an error-analysis for English. Since the profiles use

¹Available at <http://www.kyoto-project.eu/>

```

<kaf xml:lang="en" doc="example1">
  <text>
    <wf page="1" sent="40" wid="w267" fileoffset="6,11">
      water</wf>
    <wf page="1" sent="40" wid="w268" fileoffset="12,21">
      pollution</wf>
    <...>
  </text>
  <terms>
    <term id="t241" lemma="water" pos="N">
      <span><target id="w267"></target></span>
      <externalReferences>
        <externalRef conf="0.29" ref="14845743-n"
          resource="wn30g"/>
        <externalRef ref="Kyoto#water"
          reftype="sc_equivalenceOf" resource="ontology"/>
      <!--...-->
        <externalRef reftype="SubClassOf"
          reference="DOLCE-Lite.owl#endurant"
          status="implied"/>
      <!--...-->
      </externalReferences>
    </term>
    <term tid="t242" lemma="pollution" pos="N" type="open">
      <span><target id="w268"></target></span>
      <externalReferences>
        <externalRef conf="0.33" ref="14516743-n"
          resource="wn30g"/>
        <externalRef ref="Kyoto#contamination_pollution"
          reftype="sc_equivalenceOf" resource="ontology"/>
      <!--...-->
        <externalRef reftype="SubClassOf"
          reference="DOLCE-Lite.owl#perdurant"
          status="implied"/>
      <!--...-->
      </externalReferences>
    </term>
  </terms>
  <!-- Additional layers (chunking, dependencies, ...) -->
</kaf>

```

Figure 1: Example of a KAF document

language-neutral ontological constraints, they can be easily transferred to another language. Therefore in section 5, we describe how the system was transferred from English to Dutch, through the wordnet-equivalence links.

2 KYOTO overview

The KYOTO system starts with linguistic processors that apply tokenization, segmentation, morpho-syntactic analysis and semantic tagging of the text. The semantic tagging involves detection of named-entities and the meaning of words according to a given wordnet. The output of the linguistic processors is stored in an XML annotation format that is the same for all the languages, called the KYOTO Annotation Format (KAF, (Bosma et al., 2009)). KAF is compatible with the Linguistic Annotation Framework (LAF, (Ide and L.Romary, 2003)). In KAF, words, terms, constituents and syntactic dependencies are stored in separate layers with references across the structures. All modules in KYOTO draw their input from these XML structures. Likewise, WSD is done on the same KAF annotation in different languages and

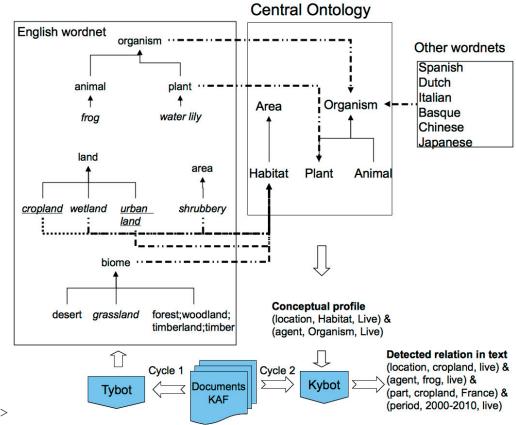


Figure 2: System Architecture

is therefore the same module for all the languages (Agirre and Soroa, 2009). The current system includes processors for English, Dutch, Italian, Spanish, Basque, Chinese and Japanese. Figure 1 shows a simplified example of a KAF structure with the two basic layers: <text> layer (tokenization, segmentation) and <terms>, containing morpho-syntactic and semantic information drawn from WordNet and the KYOTO ontology for the words *water* and *pollution*.

In KYOTO, the knowledge extraction is done by so-called Kybots (Knowledge Yielding Robots). Kybots are defined by a set of profiles representing information patterns. In the profile, conceptual relations are expressed using ontological and morpho-syntactic patterns. Since the semantics is defined through the ontology, it is possible to detect similar data even if expressed differently. In Figure 2, we show an example of a conceptual pattern for the environment domain that relates organisms that live in habitats. The pattern uses labels from the central ontology, whereas each wordnet synsets is directly or indirectly related to these labels. Such a pattern can be used to extract events from text, such as *frogs that live in cropland in France during the period 2000-2010*.

The system exploits a 3-layered knowledge-architecture (Vossen and Rigau, 2010), using a central ontology, wordnets in different languages and potential background vocabularies linked to the wordnets. The ontology consists of around 2,000 classes divided over three layers (Hicks and Herold, 2009). The top layer is based on DOLCE²

²DOLCE-Lite-Plus version 3.9.7

(Gangemi et al., 2003) and OntoWordNet. The second layer are the Base Concepts³ (BCs) which cover an intermediate level of abstraction for all nominal and verbal WordNet synsets (Izquierdo et al., 2007). Examples of BCs are: *building, vehicle, animal, plant, change, move, size, weight*. A third layer consists of domain classes introduced for detecting events and qualities in a particular domain (i.e. environment).

The semantic model also provides complete mappings to the ontology for all nominal, verbal and adjectival WordNet3.0 synsets (Fellbaum, 1998)⁴. The mappings also harmonize predicate information across different part-of-speech (POS). For instance, migratory events represented by different synsets of the verb *migrate*, the noun *migration* or the adjective *migratory* inherit the same ontological information corresponding to the *ChangeOfResidence* class in the ontology.

This generic knowledge model provides an extremely powerful basis for semantic processing in any domain. Furthermore, through the equivalence relations of wordnets in other languages to the English WordNet, this semantic framework can also be applied to other languages as shown in Section 5.

The WSD module assigns concepts to each word with a score based on the context. Ontological tagging of the text is then the last step in the pre-processing before the extraction of events. For each synset associated to a word, we use the wordnet to ontology mappings to look up its associated ontological classes and inherited properties. The Base Concept mapping guarantees that every synset is mapped to an ontological class. Next, we insert into KAF all the ontological implications that apply to each concept. By making the implicit ontological statements explicit, Kybots are able to find the same relations hidden in different expressions with different surface realizations, e.g.: *water pollution, polluted water, pollution of water, water that is polluted* directly or indirectly express the same relations. Figure 1 shows how ontological statements are represented in KAF as external references related to synsets with a score from the WSD (the value of the attribute *conf*). Words in the term structure usually get many ontological implications for each word meaning. The implications reflect subclass relations from the ontology

³<http://adimen.si.ehu.es/web/BLC>

⁴This knowledge model is freely available through the KYOTO website as open-source data.

but also other relations such as events in which the concept denoted by the word plays a role (e.g. the word *polluted water* denotes *water* that plays a role in the event *pollution*) or, the other way around, the roles involved in the event denoted by the word (e.g. the word *water pollution* denotes events in which *water* is as a patient).

3 Event extraction

A set of abstract patterns called Kybots use the central ontology to extract actual concept instances and relations from KAF documents. Event-mining is done by processing these abstract patterns on the enriched documents. These patterns are defined in a declarative format using profiles, which describe general morpho-syntactic and semantic conditions on sequences of KAF terms (which are lemmas in the text). These profiles are compiled to XQueries to efficiently scan over KAF documents uploaded into an XML database. These patterns extract the relevant information from each match.

Figure 3 shows an example of a simple Kybot profile. Profiles are described using XML syntax and consist of three main parts:

- Variable declaration (<variables> element). In this part, the search entities are defined, e.g.: **X** (terms whose part-of-speech is noun and whose lemma is not “system”), **Y** (terms whose lemma is either “release”, “produce” or “generate”) and **Z** (terms linked to a subclass of the ontological class *DOLCE-Lite.owl#contamination_pollution*, meaning *being contaminated with harmful substances*).
- Relations among variables (<rel> element): This part specifies the relations among the previously defined variables, e.g.: **Y** is the main pivot, variable **X** must precede variable **Y** in the same sentence, and variable **Z** must follow variable **Y**. Thus, this relation declares patterns like '**X** → **Y** → **Z**' in a sentence.
- Output template: describes the output to be produced for every match, e.g.: each match generates a new event from the term **Y** and two roles: the ‘done-by’ role filled by term **X** and ‘patient’ role, filled by **Z**.

We created 261 generic profiles for English. These profiles capture very simple sequences of parts-of-speech or words, e.g. noun-verb or

```

<kprofile>
<variables>
<var name="x" type="term" pos="N" lemma="! system"/>
<var name="y" type="term"
    lemma="produce | generate | release"/>
<var name="z" type="term"
    ref="DOLCE-Lite.owl#contamination_pollution"
    rtype="SubClassOf"/>
</variables>
<relations>
<root span="y"/>
<rel span="x" pivot="y" direction="preceding"/>
<rel span="z" pivot="y" direction="following"/>
</relations>
<events>
<event target="$y/@tid" lemma="$y/@lemma"
    pos="$y/@pos"/>
<role target="$x/@tid" rtype="done-by"
    lemma="$x/@lemma"/>
<role target="$z/@tid" rtype="patient"
    lemma="$z/@lemma"/>
</events>
</kprofile>

```

Figure 3: Example of a Kybot profile

adjective-noun, where each word is restricted to classes from the ontology, e.g. a motion event followed by a geographical region. Note that it is important that all possible expressions of relation are modeled by the profiles.

4 Evaluation

4.1 Triplet representation for representing events

The event structure in KYOTO is rather specific and events can be complex, including many different roles and relations. Below is an example of such a structure extracted from the sentence: “Forests also absorb air pollution and retain up to 85 percent of the nitrogen from sources such as automobiles and power plants.”

```

<event eid="e203" target="t4260"
    lemma="absorb" pos="V"
    synset="eng-30-01539633-v" rank="0.25"/>
<role rid="r280" event="e203" target="t4258"
    lemma="forest" pos="N" rtype="done-by"
    synset="eng-30-09284015-n" rank="0.15"/>
<role rid="r976" event="e203" target="t4262mw"
    lemma="air pollution" pos="N" rtype="patient"
    synset="eng-30-14517412-n" rank="1"/>
<role rid="r1609" event="e203" target="t4277mw"
    lemma="power plant" pos="N" rtype="simple-cause-of"
    synset="eng-30-03996655-n" rank="1"/>
<role rid="r276" event="e203" target="t4274"
    lemma="automobile" pos="N" rtype="simple-cause-of"
    synset="eng-30-02958343-n" rank="1"/>

```

To be able to compare our results with the output of other systems and gold-standards, we defined a more neutral and simple triplet format.

A triplet consists of:

- a relation
- a list of text token ids that represent the event

- a list of text token ids that represent a participant

If an event has multiple participants, a separate triplet is created for each event-participant pair. The triplet identifier is used to mark which triplets relate to the same event.

4.2 Evaluation results for English

We created a gold-standard in the triplet format. An annotation tool that reads KAF and can assign any set of tags to tokens in KAF was used to make a gold-standard for a document about the Chesapeake Bay, a large estuary in the US⁵. The document has 16,145 word tokens. We manually annotated all relevant relations in 127 sentences, corresponding to 1,416 tokens, 353 triplets and 201 events.

The first column in Table 2 show the annotated relations.⁶ The patient relation is most frequent (38%), followed by done-by (15%) and simple-cause-of (14%).

As a baseline, we created triplets for all heads of constituents in a single sentence according to the constituent representation of the text in KAF. The baseline generates 3,427 triplets for the annotated sentences. Since there is no relation predicted, we assume the most-frequent patient relation.

To evaluate the Kybots, we used the 261 generic profiles. The profiles generated 548 triplets for the annotated sentences. In total 169 profiles or combinations of profiles (since multiple profiles can propose the same triplet) have been applied to the annotated fragment.

To measure the proportion of relevant events that are detected by these heuristics, we compare the baseline and Kybot events with the event tokens in the gold standard. The gold standard has 201 events and the baseline 1,627 events, of which 249 overlap with the gold standard events. This results in a recall of 1.24 and a precision of 0.15. The Kybot profiles detect 733 events, of which 209 are relevant. Recall is 1.04 and precision is 0.29. Recall of events is similar to the baseline and precision is almost twice as high. The fact that the recall is higher than 1 is caused by the fact that the gold standard sometimes marks larger phrases as a single event which may be separate events in

⁵The tool and evaluation data is available at the KYOTO website

⁶The relations are taken from the DOLCE part of the KYOTO ontology. Please consult the DOLCE ontology for their formal definition.

the baseline and the Kybot output. Both the baseline and the Kybot profiles thus do not miss any relevant events but do extract a substantial amount of irrelevant events. The precision for the profiles is still reasonable, given the fact that no relevance ranking has been applied and only generic profiles have been used. Note that events that are not annotated can still be proper events.

Table 1 shows the results for the triplet evaluation of the relevant events.

	Ignoring relations		With relations	
	Baseline	Kybots	Baseline	Kybots
Nr. correct	306	222	115	174
Precision	0.09	0.49	0.03	0.32
Recall	0.86	0.63	0.33	0.49

Table 1: Baseline and Kybot results

When ignoring the relation, recall for the baseline is 86%, which shows that the baseline matches a substantial part of the annotated triplets. It also shows that 14% is missed. This is due to the fact that the parser only marks one word as the head in the case of a coordination of heads, e.g. in the phrase “birds and fish” only “bird” is marked as the head. Precision of the baseline is very low, even when we ignore the relation itself. If we take the patient relation as the default, we see that the precision and recall drop even more. The Kybot profiles clearly outperform the baseline in terms of precision: 49% when ignoring the relation and 32% considering all relations. In terms of recall, we see that 63% is covered when we ignore the relation. This indicates that the profiles do consider the majority of structures, but still miss 37% of the structures. When we consider the relations, recall drops to 49% which is still well above the baseline.

4.2.1 Error analysis

We did a separate error analysis for recall and precision. First of all, we checked the 1,023 term tokens of content words (nouns, verbs and adjectives) that occurred in the 127 gold-standard sentences. It turned out that there are 70 tokens with the wrong POS assigned (7%). The major errors are nouns and verbs interpreted as adjectives and common nouns considered as proper names, most notably “wetlands” and “wastewater” occurring 3 and 5 times respectively. If the wrong POS is assigned, the words cannot be found in WordNet or the wrong synsets are assigned. In that case, wrong or no ontological statements are inserted for a word.

To analyze the recall in more detail, we looked at the most-frequent missed relations: patient (48) and done-by (30) (see Table 2).

Relation	Gold	%	System	Correct	R.	P.	Missed
destination-of	27	7.65%	17	6	22	35	21
use-of	4	1.13%	1	1	25	100	3
generic-location	11	3.12%	22	8	72	36	3
source-of	4	1.13%	10	1	25	10	3
instrument	2	0.57%	0	0	0	0	2
product-of	2	0.57%	0	0	0	0	2
part-of	1	0.28%	3	0	0	0	1
purpose-of	7	1.98%	9	3	42	33	4
patient	133	37.68%	195	85	63	43	48
path-of	1	0.28%	0	0	0	0	1
result-of	4	1.13%	7	0	0	0	4
participant	0	0.0%	3	0	0	0	0
has-state	32	9.07%	42	11	34	26	21
state-of	22	6.23%	25	11	50	44	11
done-by	52	14.73%	89	22	42	24	30
simple-cause-of	51	14.45%	125	26	50	20	25
Total	353	100%	548	174	49	31	179

Table 2: Generic processing with 261 profiles differentiated per relation

From the patient triplets, we missed 25% due to parser errors, among which wrong-POS, missed verb-particle combinations and multiwords. Another 15% of the patient triplets was not found because the parser does not provide detailed and reliable dependency information to distinguish between subjects and objects and the ontology does not distinguish sufficiently between events with participants that control the process (e.g. “to swim”) and participants that do not (e.g. “to flow”). Remarkably, only 4% of the errors are due to a missing concept in WordNet or a wrong mapping of WordNet to the ontology. Another 4% could have been found by making more profiles.

In the case of the done-by relation, 30% of the missed relations are the result of parser errors (mainly coordination of NPs and VPs in which only one is marked as the head) and another 30% because the structures of simple-cause-of and done-by are the same and the ontology does not provide sufficient information on the events to distinguish.

Precision errors are mostly caused by the fact that patient, done-by and simple-cause-of are easily confused not only by the Kybots but also by humans. The patient relation performs slightly above average precision: 43% but done-by (24%), simple-cause-of (20%) and has-state (26%) are performing below average. Especially, the simple-cause-of relation is decreasing the overall precision since it represents 125 triplets (15%). The simple-cause-of relation applies to perdurants related to other perdurants. Due to the ambiguity in English of nouns to denote either an endurant or

a perdurant, the system is likely to over-generate this relation. The reverse holds for the done-by relation. Both relations typically hold for the same structures such as nouns in subject position of a verb. Another common error that is related are cases such as *forest destroyed* and *houses built*. Since the parser does not provide information on the inflection of the verbs nor on passive/active form, the profile can only detect a noun+verb pattern and assigns a done-by relation where a patient relation should be assigned. Again, more information from the parser in the KAF representation can help here.

The main conclusion is that major improvement both in recall and precision can be achieved by better and more input from the linguistic pre-processing, by richer ontological information e.g. control of events, and by extending the number of profiles. Furthermore, precision could also be improved if we can resolve ambiguity between endurants and perdurants of nouns to distinguish for example done-by from simple-cause-of. It thus makes sense to consider the effect of WSD on the precision of the mining. This is discussed in the next section.

4.2.2 Effects of Word-Sense-Disambiguation

The generic processing considers all the possible meanings of the words and does not take the WSD into account.

To see the effect of the WSD, we implemented a filter on the Kybot output that selects interpretations with the highest WSD score for each word in the output that has multiple interpretations. By interpretation we mean: being either an event or a role or having different relations assigned. By excluding low scoring concepts only when there is a choice to be made, we hope to capture as much recall as possible and to gain precision. Note that the WSD scored a precision of 48% in the SemEval2010 task on domain specific WSD, which used documents from the same domain as KYOTO ((Agirre et al., 2010)). We set a threshold for eliminating relations in proportion to the maximum WSD scores of each word. The results are shown in Table 3. A threshold of 0 means that all interpretations are considered, a threshold of 100 means only the highest scoring interpretations.

We can see that there is a positive correlation between WSD threshold and precision, where precision increases from 32% to 39% using the highest WSD scores only. Recall drops from 49% to

WSD threshold	#triplets	#correct	P.	R.	F1
0	548	174	0.32	0.49	0.39
10	500	169	0.34	0.48	0.40
20	479	167	0.35	0.47	0.40
30	470	167	0.36	0.47	0.41
40	461	166	0.36	0.47	0.41
50	446	164	0.37	0.46	0.41
60	434	164	0.38	0.46	0.42
70	429	162	0.38	0.46	0.41
80	427	161	0.38	0.46	0.41
90	426	161	0.38	0.46	0.41
100	377	148	0.39	0.42	0.41
manual	364	141	0.39	0.40	0.39

Table 3: Generic processing with different WSD thresholds.

42%. We get the optimal settings using a threshold for WSD of 60%. This gives an F-measure of 42%, for precision 38% and recall 46%.

We also applied a manual disambiguation of the benchmark file. The results for the manually disambiguated file are shown in the last row. We can see that less triplets are generated (364) but close to the 100% WSD threshold (377). Remarkably, the precision is the same as for 100% WSD while recall is a bit less (40%). This shows that the errors of WSD apparently do not have a big impact on the extraction. For recall, it is thus better to use less perfect WSD. This is inline with the error analysis in the previous section, which showed that structural processing is more a problem than the mapping of the text to concepts.

We also checked the effect of WSD on the extraction of relevant events. Eliminating synsets through WSD did not show any effect. Precision remains the same (29%), and recall only drops slightly from 104% to 97% when we limit the events to 100% WSD threshold. In the case of manual WSD, we do get a much higher precision (49%) and a bit lower recall (83%). This clearly shows that the Kybots over-generate many events due to the event-object ambiguity of words in English. The fact that precision of the manually tagged file is much less than the recall, also suggest that relevance of extracted events is not considered by our system: even after manual (perfect) WSD, the system detects events that are not annotated. If we consider 49% of event detection as the upper limit here, which seems reasonable, we can say that the Kybots reach a precision of 60% of the upper limit in detecting events.

4.2.3 Effects of selecting best performing profiles

The profiles perform very differently in terms of recall and precision. We therefore derived the

precision for each each profile, using the optimal WSD setting of 60% of the maximum score. We implemented a filter that checks every conflict across triplets. If two triplets involve the same events and roles but have a different relation, we choose the triplet generated by the higher scoring profile. The results are shown in Table 4. The first row of the table shows the results for a WSD threshold of 60% using 128 profiles. The remaining rows show the results when this 60% WSD output is post-filtered using profiles with precision scores 1, 5, 10, 25, 50 and 75.

	#profiles	#triplets	#correct	P.	R.	F1
All profiles	129	434	164	0.38	0.46	0.42
profiles 1%	104	332	147	0.44	0.42	0.43
profiles 5%	103	312	147	0.47	0.42	0.44
profiles 10%	103	312	147	0.47	0.42	0.44
profiles 25%	93	284	141	0.50	0.40	0.44
profiles 50%	76	219	115	0.53	0.33	0.40
profiles 75%	22	46	32	0.70	0.09	0.16

Table 4: Generic processing with WSD threshold of 60% and using best performing profiles.

We see a clear increase in precision and a drop in recall, as expected. However, we also see an increase in the F-measure from 41% to 44% using a subset of the profiles with higher precision. Using profiles with a precision score of at least 25%, we obtain a precision of 50% and a recall of 40%. With these settings, 90 profiles have been used compared to 129 profiles using just the WSD threshold of 60%. This shows that the set of profiles can be optimized for specific document collections by annotating a proportion of the collection that is representative and deriving a precision score for the different profiles. Likewise, we can pair style of writing to the type of relation expressed.

If we compare these results with the manually annotated file in Table 3, we see that the best profiles have a much higher precision (50% against 39% manual) and the same recall. This again confirms that the challenge for getting more precision is in resolving the structural relations in the text rather than assigning better concepts through WSD.

5 Transferring Kybots to another language

An important aspect of the KYOTO system is the sharing of the central ontology and the possibility to extract semantic relations in different languages in a uniform way. To test the feasibility of

sharing the same semantic backbone and transferring Kybot profiles, we carried out a transfer experiment from English to Dutch. We collected 93 Dutch documents on a Dutch estuary (the Westerschelde) and related topics. We created KAF files and applied WSD to these KAF file using the Dutch wordnet data.

To apply the profiles to the Dutch KAF documents, we need to apply the ontotagger program to the Dutch KAF. We created tables that match every Dutch synset to the English Base Concepts and to the ontology using the equivalence relations. We generated 145,189 Dutch synset to English Base Concept mappings (for comparison for English we have 114,477 mappings) and 326,667 Dutch synset to ontology mappings (186,383 for English). These ontotag tables were used to insert the ontological implications into the Dutch KAF files.

Next, we adapted the 261 English Kybot profiles to replace all English specific elements by Dutch. This mainly involved:

- replacing English prepositions and relative clause complementizers by Dutch equivalents;
- adapting the word order sequences for relative clauses in Dutch;
- adapting profiles that include adverbials, since they occur in different positions in Dutch;
- eliminating profiles for multiword compounds which mostly occur in Dutch as a one word compound;
- eliminating profiles for explicit English structures that express causal relations;

We kept all the ontological constraints exactly as they were for English. Only superficial syntactic properties were thus changed. It took us half-a-day to adapt the profiles for Dutch. From the original 261 English profiles, we obtained 134 Dutch profiles.

We ran the profiles on the 93 Dutch KAF files (42,697 word tokens) and 65 profiles generated output: 4,095 events and 6,862 roles. In terms of relations, we see a similar distribution as for English, as shown in Table 5. The patient relation is most frequent, followed by relations such as generic-location, has-state and done-by. We did a

Relation	#	%
destination-of	10	0,15%
patient	2067	30,12%
path-of	23	0,34%
has-state	1236	18,01%
generic-location	396	5,77%
state-of	748	10,90%
source-of	669	9,75%
done-by	792	11,54%
part-of	87	1,27%
simple-cause-of	573	8,35%
purpose-of	261	3,80%
Total	6.862	

Table 5: Relations extracted for Dutch documents.

preliminary inspection and the results look reasonable. For instance, two frequent words denoting events (the noun *toename* (increase) and the verb *stijgen* (increase)) appear to have sensible patients (*number, activity, consumption, pollution, trade, pressure, ground sea level, earth*).

6 Conclusions

We described an open platform for event-mining using wordnets and a central ontology that aims at maximizing the information extracted from text. The system uses a limited set of generic patterns with structural and ontological constraints on elements from the text. We have shown that wordnets can be used to map text to ontological classes and extract events and participants from text. Our error analysis showed that recall is mostly hampered by the structural complexity of the text and the incapability of the parser to handle this phenomenon. The knowledge resources, wordnet and the ontology, did not play a major role in recall. However, precision of the event relations is more affected by richness and quality of the semantics analysis. We have shown that WSD has a positive effect on the precision of the extracted relations and that precision can be further optimized by tuning the structural profiles to the genre of the target text. The system can be easily transferred to any language that has a wordnet connected to the English WordNet, as was shown for Dutch. In the future, we want to further improve recall and precision using richer event data and machine learning techniques and use the output for reconstruction of relations between events. We will also experiment with other parsers for English and Dutch to see the effect on the quality.

Acknowledgments

Partial support provided by KYOTO ICT-2007-211423 and KNOW2 TIN2009-14715-C04-04.

References

- E. Agirre E., O. Lopez de Lacalle, C. Fellbaum, S. Hsieh, M. Tesconi, M. Monachini, P. Vossen, R. Segers 2010. SemEval-2010 Task 17: All-words Word Sense Disambiguation on a Specific Domain ACL2010 workshop, July 11-16, 2010, Uppsala, Sweden, p. 75-80, Ed. K. Erk & C. Strapparava,, Publ. The Association for Computational Linguistics (ACL), ISBN 978-1-932432-70-1
- E. Agirre and A. Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *EACL*, pages 33–41.
- C. F. Baker, C. J. Fillmore, and B. Cronin. 2003. The structure of the framenet database. *International Journal of Lexicography*, 16(3):1–16.
- W. E. Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. 2009. Kaf: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation*, Pisa, Italy.
- C. Fellbaum. 1998. *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA.
- A. Gangemi, N. Guarino, C. Masolo, and A. Oltramari. 2003. Sweetening wordnet with dolce. *AI Mag.*, 24(3):13–24.
- A. Hicks and A. Herold. 2009. Evaluating ontologies with rudify. In Jan L. G. Dietz, editor, *Proceedings of the 2nd International Conference on Knowledge Engineering and Ontology Development (KEOD'09)*, pages 5–12. INSTICC Press.
- N. Ide and L. Romary. 2003. Outline of the international standard linguistic annotation framework. In *Proceedings of ACL'03 Workshop on Linguistic Annotation: Getting the Model*, pages 1–5.
- R. Izquierdo, A. Suarez, and G. Rigau. 2007. Exploring the automatic selection of basic level concepts. In Galia Angelova et al., editor, *International Conference Recent Advances in Natural Language Processing*, pages 298–302, Borovets, Bulgaria.
- K. Kaiser and S. Miksch. 2005. Information extraction. a survey. Technical report, Vienna University of Technology. Institute of Software Technology and Interactive Systems.
- L. Peshkin and A. Pfeffer. 2003. Bayesian information extraction network. In *In Proc. of the 18th International Joint Conference on Artificial Intelligence*.
- G. Rigau, A. Soroa, E. Laparra, K. Fernandez, K. Gojenola, A. Casillas, A. Díaz de Ilarrazza, and P. Vossen. 2011. Deliverable d5.4: Fact miners revised. Technical report, KYOTO project.
- P. Vossen and G. Rigau. 2010. Division of semantic labor in the global wordnet grid. In *Proc. of Global WordNet Conference (GWC'2010)*. Mumbai, India.

Upgrading WordNet: a Terminological Point of View

Cristian Zanotti

Institute for Applied Linguistics, Universitat Pompeu Fabra

Roc Boronat 138, 08018 Barcelona, Spain

Jorge Vivaldi

Mercé Lorente

{cristian.zanotti, jorge.vivaldi, merce.lorente}@upf.edu

Abstract

WordNet applied to terminology is still an ongoing issue. Due to the increasing importance of WordNet for natural language processing tools (information retrieval and extraction tools), we would like to point out its upgrade with Languages for Specific Purposes. This paper aims to discuss some issues in terminological enrichment, paying attention to both culture-dependent domains and linguistic variation matters, facts that are significant when working in a multilingual environment.

1 Introduction

From its basic design of Princeton, WordNet (WN) has been developed for many other languages. As an example, we may consider the EuroWordNet and BalkaNet projects that produced WNs for several western and eastern European languages. At the same time, WN has been produced also for languages as diverse as Hindi, Catalan or Finnish among many others¹. It has been also connected to many other resources (BabelNet, SUMO ontology, etc.) as well as enlarged and improved (eXtended WordNet, Wordnet++, etc.). It has been used for a number of interesting applications ranging from information retrieval, word sense disambiguation or lexicographical applications.

In spite of this success, the application of WN to some areas of the language, as for instance Languages for Specific Purposes (LSPs), has been scarce. This paper describes and would like to open a discussion on those peculiarities which arise when WN structure deals with specific do-

mains knowledge and in particular with legal terminology. These considerations are the product of the participation of our research group to the *Multilingual Central Repository* project (MCR) (Atserias *et al.*, 2003). This platform starts from WN 3.0 version, which serves as referring structure for the MCR; and thus, it implies WN translation into the main Iberian languages (Spanish, Catalan, Basque and Galician). For this reason, various universities around Spain have been involved in this project; in particular, our research group had the tasks to enlarge the MCR (therefore WN) with the terminology of Computer science, Environment and Law domains in both Spanish and Catalan. For this reason, this project stands out for the terminological implementation of a non-specialised database of general language. In addition, even though this paper draws its attention to legal terminology, some of the problems pointed out are also applicable to any other domain. Moreover, our collaboration to this project has, as secondary purpose, to improve the performance of YATE².

Broadly speaking, after examining the MCR, we would like to emphasise some problems which are the product of the relation between different language systems in different LSPs (Cabré, 2003). It is possible to summarise these issues as follows:

- Terminological level. It refers to how a conceptual structure, for a specific domain, is represented. At this point it is important to distinguish between:
 - Culture-independent domains: those concepts which are mostly valid even though changing cultural system; e.g.: Medicine, Technology, Computer science and Environment;
 - Culture-dependent domains: every cultural system has developed a different conceptual

¹ See: http://www.globalwordnet.org/gwa/wordnet_table.html

² Yet Another Term Extractor, described in Vivaldi (2001) and Vivaldi and Rodríguez (2002).

- structure for specific domains, such as: Law and, in part, Economics (see section 3);
- Linguistic structure. Given that WN is a lexical resource, then each language uses its own linguistic code that may differ from others.
- WN structure. Multilingual WNs are characterised by a rigid structure, basically a one to one correspondence (see section 2).

Finally, the object of this study is the terminological implementation of the MCR (a general language database). The choice for a general tool and not a specialised one can be explained as follows. Since WN provides a copy of lexical relations of the language, and being these the key element for applied linguistics applications, then it is important to improve its accuracy. Terminological development plays an important role as it concerns with two basic values: the level of precision is higher and thus it affects lexical relations; and, as a consequence, the hierarchy of a domain is described in much more depth. We believe, then, that this option would provide a more comprehensive tool for applications such as information retrieval and automatic translation.

In order to present these aspects, this paper will go into the following topics. Section 2 describes WN current status in terms of terminology and adjective treatment, pointing out that at a terminological level, adjectives are not fully represented. Then, the third section introduces the practical framework to which our work refers, drawing attention to the legal language and its peculiarities, in attempt to analyse some specific problems of translation. These issues are described in the fourth part in order to point out the difficulty to match up domains across languages and in particular two legal families inside WN. Finally, the fifth section deals with the terminological enrichment. Here the attention is centred in the implementation of relational adjectives, due to their importance for the inclusion of terminological information in WN and its usage by term extraction tools.

2 Current status

Broadly speaking, the outline of how WN has been developed since its appearance is quite complex. The WN for some languages (as those involved in the EuroWordNet project) has been developed by keeping the English WN structure and translating each of its variants. Different aspects have been taken into account for both translation and enrichment. So, in order to make

clear our context of application, we are going to briefly describe those points dealing with term and adjectives treatment aforementioned as well as some aspect in upgrading WN with terminological data.

2.1 Terminological data

First, WN was created as a general language database, afterwards, small attention was also given to the enrichment with some specific domains (Navigli and Velardi, 2002; Pociello *et al.*, 2008); such as Medicine (Gangemi *et al.*, 1999), Art and Architecture (Getty Research Institute), Geography (Buscaldi and Rosso, 2009). These are examples of domains whose concepts are not cultural dependent and thus valid for every language.

At the same time, a few examples are representative for the legal domain. First, JurWordNet (Sagri *et al.*, 2003) refers to the Italian legal domain only, so it deals with one language and one legal system. This fact do not implies particular problems of structure equivalences. Second, (Breau and Gordon, 2011) consider the problems deriving from the treatment of multi-jurisdiction and provide a way to manage it by the representation of legal documents according to a specification language.

It is clear that the biggest problems in the legal domain result when dealing with different languages and/or different legal families; these elements are vital for the development and accuracy of natural language processing tools in the domain of law (Peruginelli, 2007a). Then, in the following sections, we are going to describe some of the problems we have faced in the enrichment task.

Leaving aside these peculiarities, it is important to emphasise some main issues to consider while facing terminological data (Smith and Fellbaum, 2004):

- WN has been compiled (and often maintained/updated) by non-experts.
- Domains coverage is sparse and arbitrary.
- No domain's ontology is applied.
- Links other than hierarchical are almost ignored.

It is evident that these points are critical because they determine the accuracy of the tools built on this resource.

2.2 Adjectives in WordNet

As we are working at a terminological level, we found the need for the treatment of adjectives; due to the fact that this element provides a terminological meaning (more details in section 5.2).

As far as concerns the criteria about adjective classification and inclusion, these are described in basic WN descriptions (Fellbaum, 1998). The issues which have been highlighted refers to the adjectival implementation, disambiguation and description at a general level of the language; whereas, at a terminological level adjectives are not fully represented. More precisely, not enough attention has been given to relational adjectives, as they determine a specific feature of a term (see section 5.2). For instance, Joan *et al.* (2008) report a positive impact in the performance of YATE extractor, proved by the enlargement of a version of EuroWordNet with some relational adjectives in Economics.

Regarding the adjectival architecture of WN, polysemy analysis is one of the aspects which has been subject of interest in adjective translation and implementation (Sagot and Fiser, 2008; Soler, 2003). In the same way, another point to consider, but less represented, is when translating an element from a Source language to a Target language, thus undergoing a change of lexical category (Mohanty, 2010), due to differences in their structures. Here the author points out the necessity to organise WN at an analogue way of its speakers' knowledge.

3 Context of situation

In this section we are going to draw the attention to the legal language, as we believe it is representative of the every specific domain, as well as presenting its own peculiarities (as explained below). In fact, the majority of the matters discussed in following parts are also valid for other LSPs. For example, one of the problems we are going to discuss refers to the lack of accuracy in domains' structure. This is the case of "syndrome_2" examined in medicine. This synset is described as "has hyponym" of "symptom_2", whereas it is defined as "a combination of signs and/or symptoms that forms a distinct clinical picture indicative of a particular disorder" (CMD, 2010). Therefore, it should hang directly from "evidence_1" and share the same level of "symptom_2" and "sign_6" (MCR).

Furthermore, it is relevant to relate our study to the common status in which multilingualism is already a fact, due to globalisation of economy

and people's lives. Therefore, the current context of situation is characterised by huge transnational exchanges at both economic and social levels, and this implies the establishment of relations that go beyond individual national legal systems.

At the same time, (Peruginelli, 2007b) shows examples of multilingual and multisystem environments, such as Canada, and the European Union, where the 27 member countries have chosen to keep their own peculiarities. As a consequence, this pluralism leads us to require an increasing amount of legal information between countries of different law traditions and languages.

In this particular scenario, there are various methods and practices to manage pluralism. Regarding the discipline of Comparative law, its practical purpose (Santana, 2010) is to look at how other legal systems have solved a particular problem and which solution is closer to that situation. In fact, it is common (Pegoraro, 2010) for a High Court of a State to make use of jurisprudence from other countries. This is also explained by the fact that there are states which are more typical to export legal concepts. This is the example of legal transfers (Gillepse, 2008), which are movement of laws and institutional structures from state to state or from international organisation to state. Therefore, it is evident that legal translation plays an important role in terms of communication.

It is important, then, to understand how this requirement has a reflection on the linguistic level. It is for this reason that we are going to treat the language of law as a particular kind of LSP. In fact, legal language (Šarčević, 1997) is the product of a legal reality, which comes from how a particular territory has developed in its history, institution and culture, creating its own system of concepts. It is for this reason that, when talking about legal terminology, we refer to system-bound terms as those terms (Šarčević, 1997) that do not have a correspondence in other legal systems or families. Hence (Alcaraz, 2004), we can affirm that the main problem when translating the field of law is anisomorphism, that is the lack of conceptual symmetry between two, or more, legal systems.

Therefore, this is the problem we face the majority of the times when dealing with different legal systems; in our case, the translation task involves two different legal families: common law and civil law. The former is typical of the English language countries, while the latter is proper of continental Europe, basically Latin languages.

The results, obtained in the synsets enrichment of the legal field of the MCR project, are summarised as follows: 515 for Spanish; 578 for Catalan; 52 new Spanish, Catalan and English synsets; and 252 proposals for new adjectives.

To sum up, as concerns our point of view, it is clear that we are facing a complex situation in which it is important for professionals and academics to have tools to help their work. One of these is cross-language information retrieval, which Peruginelli (2007b) defines as “the capability for users to retrieve material written/expressed in a language different from a query language”. In LOIS project (Peruginelli, 2007b) argues that WN is a useful tool for legal information retrieval systems.

Finally, apart from this application, we observed that due to its taxonomic structure, WN has a great potential for the study of both Law and legal translation (Orozco and Sánchez-Gijón, 2011). Its hierarchical structure implies a conceptual organisation that reflects the way in which a particular domain works. In addition, the multilingual feature allows observing and comparing different legal systems. It is evident that WN structure should reflect as princely as possible the reality of legal domain, at least for those elements that are common in the majority of systems.

4 Domain structure across languages

In this section we are going to examine how the language has an effect on domain structure. This is a main issue because, as mentioned in the previous section, it is mandatory to keep WN structure across English, Spanish and Catalan.

In this task we point out two different behaviours: symmetry and asymmetry. The former refers to those concepts which have been adopted or borrowed from one country to another. This is the case for the concept *due process of law*, a legal principle which exists in Spain, but it comes from common law and it is defined as follows (MCR; Congreso de los Diputados, 2003):

ENG *due process of law*: (law) the administration of justice according to established rules and principles; based on the principle that a person cannot be deprived of life or liberty or property without appropriate legal procedures and safeguards.

The second option (asymmetry) reflects the majority of the cases. The next examples show how system-bound terms are strongly linked to the system where they have been developed. This

means that there is not a one to one conceptual correspondence between two terms, hence their conceptual delimitations do not match up exactly one to the other.

So, this situation has led us to attempt to find out an answer and, at the same time, to open a discussion on this matter. In fact, on the one hand we are facing a cultural-dependent specialised terminology and, on the other, the rigid structure of WN.

The example in Figure 1 presents a typical case of lack of correspondence, that of the lawyer profession (Cao, 2007). We observe that, at a general level, both systems share the same concept, but looking closer we notice that they cover different areas of applicability and competencies. This will result in a problem of translation, due also to the fact that WN does not provide a complete context of situation, but just an example of use. So, it is for this reason that we choose to adopt a more general term, hence not precise, to translate *lawyer* and *attorney* as *abogado*. Although, sometimes some competences (ADL, 2009) of a *solicitor* can be made by a *notario* (notary) and not an *abogado* (lawyer).

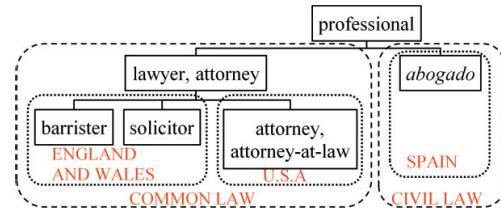


Figure 1: Lawyer profession: different conceptual delimitation in common law and civil law.

In the second example, the core of the problem lies in the variant *ticket_3*, which is defined (MCR) as: “a summons issued to an offender (especially to someone who violates a traffic regulation)”. Ticket’s definition and hierarchical structure lead us to deal with the context of traffic regulation. Since every U.S. jurisdiction has a different traffic law, then we decided to treat this issue at a general level.

For this reason, broadly speaking, we could affirm that a traffic violation is the element that makes begin the ticket legal proceeding. And therefore, it leads to the appearance before a judge, which may come to fine the offender. However, in some states, such as Maryland (MSBA, 2010), if the violation is minor, then the offender may chose to pay the fine, otherwise in

a misdemeanour case there is the obligation to appear in court.

On the contrary, from a civil law point of view, a traffic violation does not lead to begin such a legal proceeding involving a summons, but straight to a monetary sanction (*sanción pecuniaria, multa*).

Finally, we noticed that we are dealing with two different conceptual structures that correspond to two different legal proceedings; hence they do not have equivalence (Figure 3). However, in terms of legal effects we could find a correspondence between the two systems, as for instance fine and *multa*. The Figure 2 exemplifies how the two legal systems work in different ways to solve the same situation.

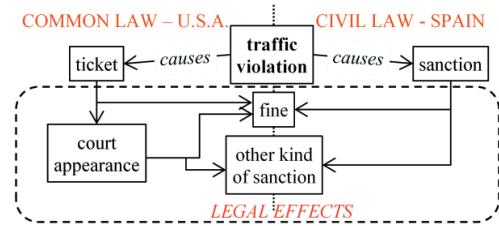


Figure 2: Exemplification of different legal proceedings due to a traffic violation.

Figure 3 represents WN current status, in which there is no chance to consider systems different from that.

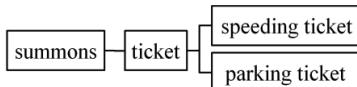


Figure 3: *Ticket* hierarchical structure in MCR.

The third example also refers to a lack of equivalence due to different conceptual systems. Common law and civil law families have developed different structure to describe and regulate criminal law. Figure 4 shows the hierarchical structure of *crime*. At first sight we notice that it does exist an equivalent for prostitution and capital offence, but it is important to point out that these do not correspond to reality, because in the Spanish legal system they are not considered as crimes. In fact, in common law countries, prostitution is not legal, while in others countries, such as Spain, only the activities which are related to the prostitution are considered as crimes.

As regards the concept *statutory offence*, this is of particular interest because it proves a lack of conceptual symmetry. In fact, it describes a

crime which is regulated by a written law (statute) and thus makes a distinction between custom, judicial precedent and statute law. On the other hand, the Spanish legal system cannot make this differentiation, because only written law is the norm.

For these reasons, providing a translation without taking into account the reality would be misleading for a member of a civil law country. Due to this fact, at this step, the choice that has been made is not to translate these terms.

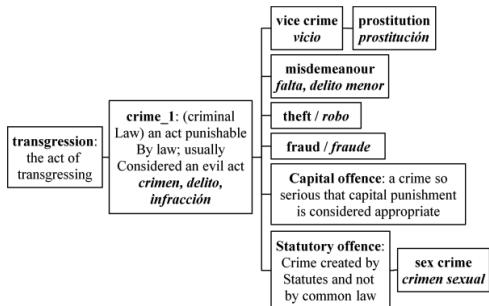


Figure 4: Criminal law: lack of conceptual equivalence.

Finally, the examples above are representative of those problems we face when dealing with legal translation, because, as we said before, they can be described as system-bound terms. So, cultural-dependent terms are crucial in the task of legal translation. Apart from this characteristic, it is important to signal that WN structure, at least for those languages that choose to keep the English WN structure, requires a one to one correspondence and this implies a difficulty in the representation of reality.

5 Enrichment task

5.1 WN structure

In the extension of the MCR original structure, one of the difficulties is to identify the exact synset where the new conceptual structure can hang from. In order to do that, a complete analysis of the concept is required. In this way, it would be easier to understand where the right place is in the synset hierarchy. And, at the same time, it is necessary the examination of the variants already existing.

Surprisingly, while doing such a task, we realised that some specific areas are very detailed in their layout, while others are not even men-

tioned³. This fact let us think that, probably, on the one hand, terms from American culture are more represented than those which are true worldwide, hence not cultural-dependent; and on the other hand, it might be the result of non-domain specialist participation in WN development. This is the case, for example, of the lack of basic concepts such as the distinction between *legal person* and *natural person*, a basic differentiation in both civil and common law (BLD, 2004).

As far as concerns the implementation with legal and natural persons, in the decision process it has been taken into account the study of the two concepts and the creation of the definition; second, the identification of both concepts as legal subject; and third, the decision to hang it from the synset entity_1, because after the comparison between the other variants, we consider that legal_subject_1 can be understood as a kind of entity_1.

Figure 5 shows a proposal for the extension with the hierarchical structure of legal subject.

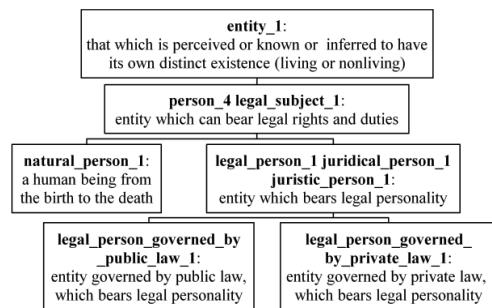


Figure 5: Proposal for the extension with legal subject and its hyponyms.

5.2 Adjectives discussion

Adjectives relevance in terminology

The second part of the enrichment refers to relational adjectives. Our starting point is that relational adjectives play an important role in the representation and communication of specialised knowledge. In fact (Estopà *et al.*, 2002), not only nominal lexical units represent knots of specialised knowledge, but at the same time, also adjectival units acquire a specialised characteristic, because they modify the nouns which are related

³ This granularity issue has been largely reported in several NLP areas (e.g., word sense disambiguation).

to. For this reason we consider the implementation of MCR adjectival structure to be relevant. In addition, thanks to this task, also term extraction would be more precise.

Furthermore, as refers to relational adjective, according to the glossary of WN (Princeton University, 2010), are usually defined as “of or pertaining to” a noun and do not have antonyms. So, for example, *legal* is a relational adjective which pertains or is related to the *law*. In addition, the example of the nominal syntagm *legal framework* clearly shows that it refers to the domain of law, while *framework* alone refers to general language.

Adjective implementation issues

Two different problems emerged while working on the adjectival implementation. The first refers to linguistic differences, which do not have a reflection on WN structure, because they are not present in it. As we are facing the translation between two languages that belong to different families (Germanic and Latin), they present clear distinction at syntactic level. This two language families has chosen different syntactic constructions. Broadly speaking, English adjectival structure can be described as follows:

- Noun + Noun
- Adjective + Noun
- Noun *OF* Noun

On the other hand, Spanish, and other Latin languages, prefer in such cases the Noun + Adjective structure. As a consequence, this is not only a different location of linguistic units, but also a different grammatical category between the two languages. This means that in most cases the Spanish relational adjective corresponds to an English noun. The following example points out the existence of the difference at grammatical level between English and Spanish, in fact what in the first language is a Noun, in the second is an Adjective:

Ind. article	Noun	Noun
An	arbitration	hearing
↓		
Ind. article	Noun	Adjective
Un	juicio	conciliatorio

It is clear, then, that this peculiarity implies an obstacle when enriching MCR with Spanish relational adjectives, and also other Latin languages,

because in most cases they do not correspond to the same English grammatical category. For this reason, it would be desirable that WN internal structure and interface would reflect the reality of facts as much as possible.

The following examples, in Table 1, are an extract of some relational adjectives which are not present in the MCR because of the syntactic

problems already mentioned. In addition, the Table 2 shows some examples which already exist in the MCR, and thus do not represent a syntactic problem. Hence, this issue would like to point out that also linguistic systems are vital when dealing with such a tool. In particular, we have observed the difficulty to represent this feature inside WN structure.

SPA adjective	Definition	Example	Relation to WN synset
<i>mayoritario</i>	<i>Pertenece o es relativo a la mayoría / Pertains or is related to the majority</i>	<i>Un acuerdo mayoritario / A majority agreement</i>	mayoría_1
<i>minoritario</i>	<i>Pertenece o es relativo a la minoría / Pertains or is related to the minority</i>	<i>Un accionista minoritario / A minority shareholder</i>	minoría_1
<i>inmobiliario</i>	<i>Pertenece o es relativo a bienes inmuebles / Pertains or is related to the real property</i>	<i>El mercado inmobiliario / The real estate market The real property market</i>	inmueble_1
<i>mobiliario</i>	<i>Pertenece o es relativo a bienes muebles / Pertains or is related to the personal property</i>	<i>El capital mobiliario / The personal property capital</i>	mueble_1
<i>pericial</i>	<i>Pertenece o es relativo a la pericia / Pertains or is related to the expert report</i>	<i>El informe pericial / The expert report</i>	peritaje [new synset]
<i>indemnizatorio</i>	<i>Pertenece o es relativo a la indemnización / Pertains or is related to the indemnification</i>	<i>Un procedimiento indemnizatorio / An indemnification proceeding</i>	indemnización_1
<i>abjuratorio</i>	<i>Pertenece o es relativo a la abjuración / Pertains or is related to the abjuration</i>	<i>Un documento abjuratorio / An abjuration document</i>	abjuración_3
<i>conciliatorio</i>	<i>Pertenece o es relativo a la conciliación / Pertains or is related to conciliation</i>	<i>Un juicio conciliatorio / An arbitration hearing</i>	conciliación_1
<i>cameral</i>	<i>Pertenece o es relativo a la cámara institucional / Pertains or is related to the chamber</i>	<i>El estatuto jurídico cameral / The chamber legal status</i>	cámara_5 cámara_7
<i>crediticio</i>	<i>Pertenece o es relativo al crédito / Pertains or is related to the credit</i>	<i>El mercado crediticio / The credit market</i>	crédito_3

Table 1: A sample of relational adjectives not existing in the MCR.

Adjective	Definition	Example	Relation to WN synset
<i>computational / computacional</i>	of or involving computation or computers	computational linguistics	computer_1
<i>technological / tecnológico</i>	of or relating to a practical subject that is organized according to scientific principles	technological development	technology_2
<i>cryptographic / criptográfico</i>	of or relating to cryptanalysis	MISSING	cryptography_1

Table 2: A sample of relational adjectives present in the MCR.

6 Conclusions

The paper has highlighted those characteristics proper of LSPs, in particular for legal terminology, especially for some problems that are the product of a multilingual structure with English as the reference language.

The examples provided are meant to demonstrate the existence of lack of symmetry at both levels of conceptual and linguistic systems. It is relevant to point out that these mismatching affect the accuracy of the resource we are upgrading. Also, we may add that this is not the product of linguistic problems but simply a reflection of the reality, as each language sets autonomously its concepts, categories, relations, variations and type of discourse and not the other way around. In fact, language naturally reflects different systems of social and community organisation.

Thus, no correspondence is the norm, at least for cultural-dependent discipline. The difficulty lies in the attempt to represent the lexical knowledge (formal, semantic and pragmatic) in a unique structure for every language. From engineers point of view would be handy to have symmetry in interlinguistic structure, but this do not correspond to reality.

At present, the MCR has a few resources to answer this problem (like a “no lexicalise” checkbox). On the contrary, we argue that this solution is not the one who best represents reality. In fact, there are cases (as described in section 4) in which a term in a second language does not pertain to the same conceptual structure of the first. At the same time, it is also important to signal that in WN enrichment we only have the choice to add hyperonymyc relations; it is nevertheless significant to make use of every existing lexical relation, in order to provide a more comprehensive resource.

Improvements may be possible in the prospect of collaboration between knowledge representation engineers, terminologists and domain experts. By doing this, a resource such as WN will continue to be useful in automatic translation, information retrieval and knowledge management, without betraying linguistic variation.

On the other hand, even though WN presents *defects* (Smith and Fellbaum, 2004), still, “it is the principle lexical database used in natural language processing research and applications”. The MCR takes as reference the English WN and is characterised by a domains’ ontology more pre-

cise (Agirre *et al.*, 2007) than WN 2.0 (Smith and Fellbaum, 2004), besides it presents a multi-language and multi-domain environment.

The advantages for a database that covers different sciences lie in its applications. In fact, tools such as term extractors will work more accurately for the recognition of which domain a term pertains to. Otherwise, the use of a single domain database would not produce the same performance, or it would need to be completed by other general language resources.

In conclusion, we first observed WN structure at both terminological and adjectival levels, pointing out its peculiarities in a multilingual environment. The reasons for a multilingual tool are motivated explaining that current human relations imply a continuous exchange of information between different languages and legal structures. After that, we described the issues faced when translating and enriching. In the former point, cases of conceptual asymmetry were pointed out. Whereas, in the latter, WN structure was argued: first the lack of not cultural-dependent area of knowledge and then, the question of adjectival structure matching between English and Latin languages. In our opinion, some improvements/upgrading in the structure of WN synsets seems to be necessary in order to solve the issues shown.

Finally, we would like to emphasize the fact that a multi-domain and multilanguage database is a more comprehensive tool which could help to improve those applications based on it. At the same time, it is also desirable a cooperation between experts from different scientific areas, in order to help WN architecture being more coherent in its task of representation of reality.

Acknowledgments

This research has been supported by the Science and Education Ministry (Spain) under the Rico-Term4 project (FFI2010-21365-C03-01) and the Special Action: MCR:IULA (FFI2009-08484-E/FILO).

References

- ADL, *A Dictionary of Law*. 2009. Law, J. and Elizabeth A., editors. Oxford University Press.
- Agirre, E., Alegria, I., Rigau, G., Vossen, and P. 2007. MCR for CLIR. *Procesamiento del Lenguaje Natural*, 38, abril 2007:3-15.
- Alcaraz, Enrique. 2004. Anisomorfismo y lexicografía técnica. In L. González and P. Hernández, coords,

- Las palabras del traductor.* Eslétra, European Commission, Brussels, pages 201-220.
- Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., and Vossen, P. 2003. The MEANING Multilingual Central Repository. In *Proceedings of the Second International Global WordNet Conference (GWC'04)*. Brno, Czech Republic.
- BLD, Black's Law Dictionary, eighth edition.* 2004. Bryan A. Garner, editor. Thomson business, St. Paul, MN, U.S.A.
- Breau, Travis and Gordon, David. 2011. Managing Multi-Jurisdictional Requirements in a Computational Legal Landscape. Technical Report, CMU-ISR-11-102.
- Buscaldi, D. and Rosso, P. 2009. Using GeoWordNet for Geographical Information Retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access, LNCS*, 5706(2009):863-866. ISSN 0302-9743.
- Cabré, María Teresa. 2003. Theories of terminology. Their description, prescription and explanation. *Terminology*, 9(2):163-200. John Benjamins, Amsterdam. ISSN 1569-9994.
- Cao, Deborah. 2007. *Translating Law*. Multilingual Matters, Clevedon.
- CMD, Concise Medical Dictionary, eighth edition.* 2010. Elisabeth Martin, editor. Oxford University Press.
- Congreso de los Diputados. 2003. *Sinopsis artículo 24, Constitución española*. Retrieved from: <http://www.congreso.es/consti/constitucion/indice/sinopsis/sinopsis.jsp?art=24&tipo=2>.
- DIEC, Diccionari de la llengua catalana, segona edició.* 2007. Institut d'Estudis Catalans, Barcelona. Retrieved from: <http://dlc.iec.cat/>.
- Estopà, R., Lorente, M., and Folquerà, R. 2002. El rol de los adjetivos en los textos especializados. In *RITerm La Terminología, entre la globalización y la localización. Actas del VIII Simposio Iberoamericano de Terminología*, Cartagena, Colombia [CD-ROM].
- Fellbaum, F., Gross, D., and Miller, K. 1993. Adjectives in WordNet. In *Five papers on WordNet*. Pages 26-39. Retrieved from: <http://wordnetcode.princeton.edu/5papers.pdf>.
- Fellbaum, Christiane, editor. 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.
- Gangemi, A., Pisanelli, D.M., and Steve G. 1999. Overview of the ONIONS project: Applying Ontologies to the Integration of Medical Terminologies. *Data and Knowledge Engineering*, 31(1999):183-220.
- Gillespie, John. 2008. Towards a discursive analysis of legal transfers into developing East Asia. *International Law and Politics*, 40(657):657-722.
- Joan, A., Vivaldi, J., and Lorente, M. 2008. Turning a term extractor into a new domain: first experiences. In *LREC 2008 Proceedings*, pages 748-752, Marrakech, Morocco.
- Law, J. and Elizabeth A. 2009. *A Dictionary of Law*. Oxford University Press.
- MCR, Multilingual Central Repository. Consulted via the Web EuroWorNet Interface. Retrieved from: <http://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl>.
- MSBA, Maryland State Bar Association. 2010. So you've received a traffic ticket. In *Publications*. Retrieved from: www.msba.org.
- Mohanty, Panchanan. 2010. WordNets for Indian Languages: Some Issues. In *Global WordNet Conference 2010*, Mumbai, India.
- Navigli, R. and P. Velardi. 2002. Automatic Adaptation of WordNet to Domains. In *Proceedings of the OntoLex 2002*, pages 45-53, Las Palmas, Spain.
- Orozco, Mariana and Sánchez-Gijón, Pilar. 2011. New resources for legal translators. *Perspectives: Studies in Translatology*, 19(1):25-44.
- Pegoraro, Lucio. 2008. La utilización del derecho comparado por parte de las Cortes Constitucionales: Un análisis comparado. *La Ciencia del Derecho Procesal Constitucional. Estudios en homenaje a Héctor Fix-Zamudio en sus cincuenta años como investigador del Derecho*. Instituto Mexicano de Derecho Procesal Constitucional, Mexico, pages 285-436.
- Peruginelli, Ginevra. 2007a. Multilingual legal information access: an overview. In E. Chiocchetti and Voltmer, editors, *Harmonising legal terminology*. EURAC, Bolzano, Italy, pages 6-34.
- Peruginelli, Ginevra. 2007b. Towards a common understanding of law in a Multilanguage world: The role of cross-language legal information retrieval systems. *The European Legal Forum*, 1(2), 64-71. Retrieved from www.european-legal-forum.com.
- Pociello, E., Gurrutxaga, A., Eneko Agirre, E., Aldezarbal, I., and Rigau, G. 2008. WNTERM: Enriching the MCR with a Terminological Dictionary. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, pages 1778-1784, European Language Resources Association, Marrakech, Morocco.
- Princeton University. 2010. A glossary of WordNet terms. Retrieved from: <http://wordnet.princeton.edu/man/wngloss.7WN.html>.
- Sagri, M. T., Tiscornia, D., and Bertagna F. 2003. JurWordNet. In *Proceedings of the Second Interna-*

tional Global WordNet Conference (GWC'04), pages 305-310, Brno, Czech Republic.

Sagot, Benoît, and Fišer, Darja. 2008. Building a free French wordnet from multilingual resources. In *Proceedings of OntoLex 2008*, Marrakech, Morocco.

Santana, María. 2010. El derecho comparado en la jurisprudencia del Tribunal Constitucional español. *ReDCE*, 7(14), 427-447.

Šarčević, Susan. 1997. *New approach to legal translation*. Kluwer Law, The Hague, Netherlands.

Soler, Clara. 2003. Extension of the Spanish WordNet. In *Proceedings of the Second International Global WordNet Conference (GWC'04)*, pages 213-219, Brno, Czech Republic.

Smith, Barry, and Fellbaum, Christiane. 2004. Medical WordNet: A New Methodology for the Construction and Validation of Information Resources for Consumer Health. In *Proceedings of Coling: The 20th International Conference on Computational Linguistics*, pages 31-38, Geneva.

Vivaldi, Jorge. 2001. Extracción de candidatos a término mediante combinación de estrategias heterogéneas. PhD thesis, Universitat Politècnica de Catalunya, Barcelona. [Also in Sèrie Tesis, num. 9, Institut Universitari de Lingüística Aplicada of Universitat Pompeu Fabra].

Vivaldi, J., and H. Rodríguez. 2002. Medical Term Extraction using the EWN ontology. In *Proceedings of Terminology and Knowledge Engineering*, pages 137-142, Nancy.

SENEQA – System for Quality Testing of Wordnet Data

Tomáš Čapek

Faculty of Informatics

Masaryk University

Botanická 68a, 602 00 Brno

Czech Republic

xcapek1@aurora.fi.muni.cz

Abstract

Construction of a semantic network from scratch is a long process that usually requires both linguistic work done by hand and semi-automatic methods to add or translate new data which must be subsequently reviewed by human lexicographers. In this process, many systemic and/or language-specific errors usually appear in the data over time. Maintaining content integrity and high quality of data in a general-purpose semantic network, that is in development, is of utmost importance for majority of NLP applications in which a wordnet is used. In this article, we will introduce a prototype system to tackle this issue systematically.

1 Introduction

A general-purpose semantic network is a language resource of high-density of information and represents an alternative to traditional dictionaries. It consists of semantic units which are connected by semantic relations, thus creating a graph-like structure, or a network. The biggest semantic network to date is WordNet (PWN) (Miller et al., 1990), that has been in development since 1985 at Princeton University. It contains almost 120,000 semantic units called synsets or synonymous sets. Many WordNet-like semantic networks exist today for other languages, developed in projects such as EuroWordNet (Vossen, 1999), BalkaNet (Tufis et al., 2004) or independently. The cooperative framework of EuroWordNet is continued through the Global WordNet Association and more than sixty language databases exist today including languages such as Nepali, Korean or Afrikaans (Association, 2011).

To create a semantic network requires a team of linguists, software support and months of work,

among other things. In order to save time or resources one or both methods described below are usually employed:

- Semi-automatic translation of semantic units from another, larger network. This also refers to so-called *expand model* in EuroWordNet terminology (Vossen, 1996). In this method, we adopt the original structure of semantic units and semantic relations among them, translate each lexeme automatically via an electronic dictionary or translator system available for our language, and review the data afterwards by hand. Additional language-specific and other data are subsequently added to the network, thus *expanding* it. This is generally the fastest and the most popular method when creating a new wordnet¹ semantic network and PWN is the most accessible semantic network used as a template. This method is the also most prone to adopting and creating new errors when used as the only method.
- Manual linguistic work – also called the *merge model*² in EuroWordNet terminology. The main focus here is to create a semantic network with independent structure or predetermined application in mind. An existing language resource can be used as the source lexicon, data in which need to be rearranged and interconnected via semantic relations to form a semantic network. This method is similar to traditional construction of dictionaries and is known to be very time-consuming and expensive. It also requires a lot of linguistic introspection on part of the developers and can be outsourced, so that

¹Whereas **WordNet** stands for original Princeton WordNet (PWN), **wordnet** written in lower-case letters represents any semantic network build upon principles of PWN.

many different people take turn in the process of adding and editing the data. As an implication, semantic networks built according to this method are also prone to contain many types of inconsistencies and errors.

2 In-development Integrity Control

There are several ways we can take to prevent errors from appearing in our network while it is still in development. However they bring additional expenses on time, resources and manpower. As evidenced in literature, these additional means have often not been used in practise (Sojka et al., 2003), (Sojka et al., 2005), (Tancs et al., 2007).

2.1 Corpus Evidence

When adding, checking, or translating lexemes and semantic units, it is important to have an appropriate corpus available as the definitive source of real-life usage of words. No two lexicographers have exactly the same knowledge and perspective of a language and that perspective changes even for a single lexicographer over time. In this regard, corpora help to streamline and unify otherwise divergent approaches to handle linguistic data, especially those occurring with low frequency. The bigger the corpus is, the better but it is also important for it to contain only relevant documents with respect to the contents of the semantic network. Unsorted pile of random documents can provide false or inaccurate evidence for the lexicographers, thus spoiling the benefits corpora can bring to the process of a semantic network development. Creating a custom corpus of a preselected domain might be the best way to control the quality and relevance of texts (Baroni et al., 2006).

2.2 Guideline Manual

A guideline manual, or a style guide, is a set of instructions how lexicographers should handle new or existing semantic units and relations. It sets the standard for people who participate in the semantic network development and who may come and go as the process goes on. It should provide basic information on issues such as:

²EuroWordNet was a project primarily focused to create a multi-lingual semantic network based on PWN. The *merge* part of the process refers to the final stage of development when the semantic units in the newly-created wordnet are connected to their corresponding counterparts in another network, thus *merging* it into one bilingual structure.

- what are the criteria for a word to be lexicalized or non-lexicalized in the data,
- in what way to compose or assume definitions for semantic units,
- how to use notes for further work,
- what semantic relations are important for a particular part of speech.

The nature of the guidelines should be dependent on the aim of the semantic network itself. The guidelines also provide specification for implementation into a software tool, which is used for editing of semantic data. For example, *is-a* might be set as a required semantic relation for each noun synset and the software tool would not allow to save a synset entry, unless this relation has been attached to the synset.

2.3 Quality Testing

Ideally, any data in a wordnet should be checked and reviewed by lexicographers who have not been involved in the data development, to prevent bias. As we have discussed, there are several ways how to import erroneous data into wordnet data. It has been argued that quality of semantic data is directly related to the success rate of any NLP application that employs it, or to the usefulness when used as another language resource for linguistic work (Piasecki et al., 2009). If no guidelines exist for a given semantic network then quality assurance may result in ad hoc fixes or random edits because it is not clear what aspects of development were important in the past or when they may change again. Thus, the quality testing basically means a check to what extent the semantic data conform the guidelines.

3 Heuristic Tests

As outlined above, contrary to our best intentions, many different errors and inconsistencies appear in wordnet data over time. Such errors become relevant when we need to use the data in an NLP experiment but don't have time and resources to fix the data directly. One way to quickly analyze the data is to design and implement a set of heuristic tests. A heuristic test is a formalized pattern of an error that appears multiple times in wordnet lexical data. For example, Czech orthography allows us to use two different suffixes in words ending with *-ism* (e.g. in albinism). We can use a

suffix with *s* or *z* in it – both *albinizmus* and *albinismus* are correct word forms in Czech. However, it may be useful in more than one way to use only one suffix variant consistently. In this case the test is very simple, we choose a suffix and let the test search for words with the other suffix. On the output, we get a list of candidate literals for review. The next step is to simply replace the words as there is virtually no possibility we could get a false positive from this test in Czech.

Most of semantic networks continue to be edited even after the main development project has ended. Once a test is implemented, it is useful to have it scheduled for regular runs after a certain period of time via the *cron* tool (Keller, 1999) or any other job scheduler software. The results are then automatically reported via e-mail which helps to keep the integrity of the wordnet lexical data up-to-date at all times.

4 Classification of Errors

In general, we recognize two types of errors in a wordnet – surface errors and structural errors. The distinction is based on how difficult and time-effective it is to find errors of the given category. The surface errors are directly present in lexical units, synset literals, glosses, or other metadata thereof, and when discovering such errors, we can work with word lists, ignoring the rest of the semantic network structure. We can also find certain cases of mismatched or missing lexical relations among synsets if we only need to check whether a relation exists with a synset rather than checking both ends of such a relation.

Structural errors are harder to find and sometimes also hard to define. They deal with correctness and appropriateness of lexical and semantic relations among synsets, of literals in richly-populated synsets, balance of synset subtrees, information density, or granularity of sense variants. Finding these errors requires much more effort and at the same time it exhibits lower precision when done automatically. Tests to discover and fix structural errors are not yet implemented and remain as an objection for the future. In the following, let's consider some examples of useful surface tests:

- **Morphology tests** In this category of tests we check for typing errors or for incorrect word forms, lemmata of which belong to the network. As a requirement, we need a spell checking tool and a dictionary for

our language (e.g. *ispell* (Kuenning et al., 2004)) but for highly inflectional languages such as Czech and other Slavonic languages it is far more useful to employ a morphological analyzer that can generate and recognize any word forms belonging to the language (Sedláček and Smrž, 2001), (Šmerk, 2009). If we use the *expand* model or use other means to automatically add semantic units for subsequent translation (Blahuš, 2011), (Němčík et al., 2008), morphology test can also filter out unknown, untranslated or otherwise inappropriate words.

- **Syntax tests** If no formal guidelines have been used during development of the lexical data, any type of unexpected data can get into the synset entries. Usually, they are various notes from the lexicographers or redundant characters that remained after automatic import procedures from other language resources. A simple test for non-letter characters and for high word counts in lexeme records can discover potentially erroneous semantic units. The advantage of this test is that it is much cheaper to employ than to implement a full set of syntactic restrictions directly into the software editing tool that is used to work with the data.
- **Instance test** Many cases of semantic relation pair class-instance (e.g. sea-Aegean Sea) are often marked as simple cases of hyperonymy-hyponymy in many semantic networks. To remedy this, a test that recognizes named entities is required to filter out words and collocations which should have the semantic relations to their superordinate synset changed to *Instance*.
- **Orphan nodes** Each part of speech has at least one significant semantic relation that connects all synsets of its kind. For instance it is hyperonymy-hyponymy (*is-a*) pair for nouns. When new data are added to the network by hand or automatically, some of the entries remain unconnected to any other entry, thus creating orphan nodes in the lexical data. A simple test can discover these nodes by checking each synset for that particular se-

mantic relation. If the test results are to be reviewed by hand, this test can be extended to other semantic relations as well, even if they are not supposed to interconnect every synset of the particular part of speech.

Apart from the tests above many other language-specific or general tests can be designed according to particular needs of each wordnet. It should always be quicker to implement a test, if we can find a pattern in the data, rather than to do a full revision in top-down or alphabetical order.

5 SENEQA

A collection of mutually independent heuristic tests, along with other necessary software infrastructure, became a foundation for a prototype system, named SENEQA (SEmantic NEtwork Quality Assurance). Each test connects to the lexical database in the XML format and reports entries that need to be checked or edited, and in some cases can fix the candidate entries automatically.

The architecture of SENEQA aims to provide maximum flexibility in design to create and run different tests. the way to achieve it is to exploit available low-level NLP tools, that were developed independently, as much as possible. For Czech, we employ a morphological analyzer (Sedláček and Smrž, 2001), a recognizer for named entities and multi-word expressions (Čapek and Šmerk, 2010) or word sketches (Kilgarriff et al., 2004). With similar tools available for a different language, all tests can be easily adopted and converted to be used with a different wordnet.

Currently, five tests are implemented for Czech WordNet (CZWN, currently containing 28,486 synsets and 76,009 literals) that automatically check for error patterns:

- *basic test* - detects typographical and/or spelling errors, or otherwise incorrectly entered words, non-Czech expressions and also valid Czech words unknown to morphological analyser of Czech. An error of this type has been detected in almost 3 % of all CZWN literals.
- *wrong-pos test* - detects synsets with part-of-speech tag mismatch. Results for this test are still unprocessed for CZWN.
- *false roots test* – detects synsets which lack a basic relation with another synset for its part

of speech – in other words, they miss their anchor in the semantic network structure. For example a basic relation for nouns is *is-a* relation or its equivalent. In CZWN, nearly 10 % of all synsets have been detected with this error.

- *multiple mwe test* – detects multi-word expressions in synset literals which occur multiple times in the semantic network. Such literals are most likely redundant or incorrect. 523 literals have been detected in the most recent version of CZWN exhibiting this phenomenon.
- *instance literals test* – detects synsets which represent an instance of a certain class in the network but are linked to the corresponding synset by a different type of relation, typically the *is-a* relation or its equivalent. 355 candidate synsets have been detected in the most recent version of CZWN exhibiting this phenomenon.

6 Discussion and Future Work

A list of entries produced by a heuristic test can be processed automatically though it is often necessary or advisable to review it manually. In any case, a heuristic test usually produces a list of entries (corresponding to an error pattern), which is of lower-order volume compared to the semantic network as a whole. This greatly facilitates the quality assurance workflow as each test focuses on a couple of closely related issues at most. SENEQA also allows us to strictly separate maintenance of the lexical data from its development in the following way: if at any time, such as during processing a heuristic test results, an error is discovered that is not yet covered by any test, then a new and independent test should be developed to produce a separate list of all entries in the semantic network exhibiting the same type of error. This way, the whole process is straightforward and systematic.

The SENEQA system is not meant to be a cure-all as far as present or future inconsistencies in wordnet lexicons are concerned; rather, it allows transparent incremental development. With each new test, one whole error pattern disappears from the data and should not pose a problem in the future.

Although the heuristic tests are often very simple and quick to implement they only cover the

surface errors and inconsistencies visible at first sight. They can also help us to find various structural defects in the network such as undesired multiple inheritance, unbalanced trees or high sense number count for a lexeme but cannot offer an automatic solution for such problems. Our further work with SENEQA will therefore be focused on more sophisticated methods that will allow us to tackle practical problems with ontologies, data density or domain subtrees in a semantic network.

7 Conclusion

We have discussed an issue on how to create and maintain semantic data in a wordnet semantic network that would allow us to minimize the number of errors and inconsistencies on surface level of the network. We have introduced SENEQA prototype system that uses a collection of heuristic tests to help us remove frequent errors in the data even when the wordnet is still in development and many lexicographers participate in its creation. Although the tests are not an universal remedy to all problems that might occur in the lexical data, their favorable cost-benefit ratio makes them a useful tool to keep the integrity of the data intact.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and LINDAT-Clarin project LM2010013, and by the Czech Science Foundation under the projects P401/10/0792.

References

- Global WordNet Association. 2011. Wordnets in the World. http://www.globalwordnet.org/gwa/wordnet_table.htm.
- M. Baroni, A. Kilgarriff, J. Pomíkálek, and P. Rychlý. 2006. Webbootcat: a web tool for instant corpora. In *Proceeding of the EuraLex Conference*, pages 123–132.
- Marek Blahuš. 2011. Extending Czech WordNet Using a Bilingual Dictionary. Master’s thesis, Masaryk University, Brno.
- M. S. Keller. 1999. Take command: cron: Job scheduler. *Linux Journal*.
- A. Kilgarriff, P. Rychlý, P. Smrž, and D. Tugwell. 2004. The sketch engine. In *Proceedings of EURALEX*, pages 105–116.
- G. Kuemming, P. Willisson, W. Buehring, and K. Stevens. 2004. International ispell. *Webpage can be found at: http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell.html*, visited on February 17th.
- G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, and R. Tengi. 1990. Five papers on WordNet. *International Journal of Lexicography*, 3(4):235–312.
- V. Němcík, K. Pala, and D. Hlaváčková. 2008. Semi-automatic linking of new czech synsets using princeton wordnet. In *Intelligent Information Systems XVI, Proceedings of the International IIS’08 Conference*, pages 369–374.
- Maciej Piasecki, Stanisaw Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Oficina Wydawnicza Politechniki Wrocławskiej.
- R. Sedláček and P. Smrž. 2001. A new czech morphological analyser ajka. In *Text, Speech and Dialogue*, pages 100–107. Springer.
- Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, and Piek Vossen, editors. 2003. *Proceedings of the Second International WordNet Conference—GWC 2004*, Brno, Czech Republic, December. Masaryk University Brno, Czech Republic.
- Petr Sojka, Key-Sun Choi, Christiane Fellbaum, and Piek Vossen, editors. 2005. *Proceedings of the Third International WordNet Conference—GWC 2006*, South Jeju Island, Korea. Masaryk University Brno, Czech Republic.
- Attila Tancs, Dra Cséndes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors. 2007. *Proceedings of the Fourth Global WordNet Conference—GWC 2008*, Szeged, Hungary. University of Szeged, Department of Informatics.
- D. Tufis, D. Cristea, and S. Stamou. 2004. Balkanet: Aims, methods, results and perspectives. A general overview. *SCIENCE AND TECHNOLOGY*, 7(1-2):9–43.
- T. Čapek and P. Šmerk. 2010. Towards Partial Word Sense Disambiguation Tools for Czech. In *Proceedings of the RASLAN Workshop 2010*, pages 103–108. Masaryk University, Brno.
- P. Vossen. 1996. Right or Wrong. Combining lexical resources in the EuroWordNet project. In *M. Gellerstam, J. Jarborg, S. Malmgren, K. Noren, L. Rogstrom, CR Papmehl, Proceedings of Euralex-96, Gothenburg*, pages 715–728. Citeseer.
- P. Vossen. 1999. Eurowordnet a multilingual database with lexical semantic networks. *Computational Linguistics*, 25(4).
- P. Šmerk. 2009. Fast Morphological Analysis of Czech. In *Proceedings of the RASLAN Workshop 2009*, pages 13–16. Masaryk University, Brno.

Author Index

author	page	paper
A		
A R, Balamurali	10	Leveraging Sentiment to Compute Word Similarity
Abdul-Mageed, Muhammad	18	Toward Building a Large-Scale Arabic Sentiment Lexicon
Abrate, Matteo	23	WordNet Atlas: a web application for visualizing WordNet as a zoomable map
Adhikary, Das	205	Ontology of Sanskrit Wordnet: Nouns and Verbs
Ajotikar, Tanuja	30	Verbs in Sanskrit Wordnet
Almási, Attila	377	Multiword Verbs in WordNets
Angioni, Manuela	365	A Linguistic Approach to Feature Extraction Based on a Lexical Database of the Properties of Adjectives and Adverbs.
Ansa Osteriz, Olatz	260	Using WordNet into UKB in a Question Answering System for Basque
Anwarus Salam, Khan Md.	35	Using WordNet to Handle the OOV Problem in English to Bangla Machine Translation
B		
Bacciu, Clara	23	WordNet Atlas: a web application for visualizing WordNet as a zoomable map
Baldwin, Timothy	56	Japanese SemCor: A Sense-tagged Corpus of Japanese
Ben Aouicha, Mohamed	126	New WordNet-based semantic relatedness measurement: Using new information content metric and k-means clustering algorithm
Bertorello, Anna Rita	40	A new hierarchy of ArchiWordNet (AWN): building parts implementation with image
Bharali, Himadri	155	ASSAMESE VOCABULARY AND ASSAMESE WORDNET BUILDING: AN ANALYSIS
Bhatt, Brijesh	45	Introduction to Gujarati wordnet
	10	Leveraging Sentiment to Compute Word Similarity
	30	Verbs in Sanskrit Wordnet
Bhattacharyya, Pushpak	45	Introduction to Gujarati wordnet
	79	A Study of the Sense Annotation Process: Man v/s Machine
Bhensdadia, C	45	Introduction to Gujarati wordnet
Blahuš, Marek	50	Extending CzechWordNet Using a Bilingual Dictionary
	56	Japanese SemCor: A Sense-tagged Corpus of Japanese
Bond, Francis	64	A Survey of WordNets and their Licenses
	211	Using WordNet to predict numeral classifiers in Chinese and Japanese
Borin, Lars	254	Linking and Validating Nordic and Baltic Wordnets – A Multilingual Action in META-NORD

author	page	paper
Buttler, David	338	The C-Cat Wordnet Package: An Open Source Package for modifying and applying Wordnet
C		
Castellón, Irene	72	Semantic Hand-Tagging of the SenSem Corpus Using Spanish WordNet Senses
Chatterjee, Arindam	79	A Study of the Sense Annotation Process: Man v/s Machine
Chauhan, Dinesh	45	Introduction to Gujarati wordnet
Climent, Salvador	72	Semantic Hand-Tagging of the SenSem Corpus Using Spanish WordNet Senses
	232	Building WordNets by machine translation of sense tagged corpora
Coll-Florit, Marta	72	Semantic Hand-Tagging of the SenSem Corpus Using Spanish WordNet Senses
Csirik, János	377	Multiword Verbs in WordNets
D		
Dahamna, Badisse	94	Combining Wordnet and Crosslingual multi-terminology health portal to access health information
Dai, Lin	86	A Computer Aided Approach for Enriching WordNet with Semantic Definition
Darmoni, Stefan J	94	Combining Wordnet and Crosslingual multi-terminology health portal to access health information
De Paiva, Valeria	100	Revisiting a Brazilian WordNet
Diab, Mona	18	Toward Building a Large-Scale Arabic Sentiment Lexicon
Dutoit, Dominique	94	Combining Wordnet and Crosslingual multi-terminology health portal to access health information
E		
Erjavec, Tomaž	113	sloWNet 3.0: development, extension and cleaning
F		
Fellbaum, Christiane	105	Scalar Properties of Emotion Verbs and Their Representation in WordNet
	142	Restructuring Adjectives in WordNet with ClusterEditor
	173	Rethinking WordNet's Domains
	330	Refining WordNet adjective dumbbells using intensity relations
Fišer, Darja	113	sloWNet 3.0: development, extension and cleaning
	317	Automatic Extension of WOLF
Forsberg, Markus	254	Linking and Validating Nordic and Baltic Wordnets – A Multilingual Action in META-NORD
Fothergill, Richard	56	Japanese SemCor: A Sense-tagged Corpus of Japanese

author	page	paper
Frontini, Francesca	219	Mapping a corpus-induced ontology of action verbs on ItalWordNet
G		
Gagliardi, Gloria	219	Mapping a corpus-induced ontology of action verbs on ItalWordNet
Gao, Eshley	211	Using WordNet to predict numeral classifiers in Chinese and Japanese
Gonzalez, Aitor	118	Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base
Griffon, Nicolas	94	Combining Wordnet and Crosslingual multi-terminology health portal to access health information
Grosjean, Julien	94	Combining Wordnet and Crosslingual multi-terminology health portal to access health information
Grác, Marek	299	Low-cost ontology development
H		
Hadj Taieb, Mohamed Ali	126	New WordNet-based semantic relatedness measurement: Using new information content metric and k-means clustering algorithm
Hayashi, Yoshihiko	134	Computing Cross-Lingual Synonym Set Similarity by Using Princeton Annotated Gloss Corpus
Horák, Aleš	305	Migrating Cornetto Lexicon to New XML Database Engine
Huang, Chu-Ren	294	Introducing WordNet in Interpreting Studies – Implications and Desiderata
Huang, Heyan	86	A Computer Aided Approach for Enriching WordNet with Semantic Definition
Huang, Terry	338	The C-Cat Wordnet Package: An Open Source Package for modifying and applying Wordnet
Hyvärinen, Mirka	227	Using a Bilingual Resource to Add Synonyms to a Wordnet: FinnWordNet and Wikipedia as an Example
I		
Indyka-Piasecka, Agnieszka	268	Automated Mapping of Polish Proper Names on plWordNet Hypernymy Structure
J		
Joshi, Salil	79	A Study of the Sense Annotation Process: Man v/s Machine
Julien, Isaac	142	Restructuring Adjectives in WordNet with ClusterEditor
	330	Refining WordNet adjective dumbbells using intensity relations

author	page	paper
--------	------	-------

K

- Kalita, Chandan** 149 An Extractive Approach of Text Summarization of Assamese using WordNet
Kanojia, Diptesh 79 A Study of the Sense Annotation Process: Man v/s Machine
Kawai, Atsuo 349 Extension of Phrases for Article Determination using WordNet Thesaurus
Kr. Sarma, Shikhar 155 ASSAMESE VOCABULARY AND ASSAMESE WORDNET BUILDING: AN ANALYSIS
Kulkarni, Malhar 30 VERBS IN SANSKRIT WORDNET
Kurc, Roman 268 Automated Mapping of Polish Proper Names on plWordNet Hypernymy Structure

L

- Laparra, Egoitz** 118 Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base
Lebani, Gianluca 159 Encoding Commonsense Lexical Knowledge into WordNet
Li, John 248 Wordnet and SUMO for Sentiment Analysis
227 Using a Bilingual Resource to Add Synonyms to a Wordnet: FinnWordNet and Wikipedia as an Example
Lindén, Krister 254 Linking and Validating Nordic and Baltic Wordnets – A Multilingual Action in META-NORD
286 Finding a Location for a New Word in WordNet
Lloberes, Marina 72 Semantic Hand-Tagging of the SenSem Corpus Using Spanish WordNet Senses
Lohk, Ahti 167 Visual Study of Estonian Wordnet using Bipartite Graphs and Minimal Crossing algorithm
Lorente Casafont, Mercè 390 Upgrading WordNet: a Terminological Point of View

M

- Ma, Xiaojuan** 173 Rethinking WordNet's Domains
Mahanta, Mayashree 155 ASSAMESE VOCABULARY AND ASSAMESE WORDNET BUILDING: AN ANALYSIS
Malu, Akshat 10 Leveraging Sentiment to Compute Word Similarity
Marchetti, Andrea 23 WordNet Atlas: a web application for visualizing WordNet as a zoomable map
Mathieu, Yvette Yannick 105 Scalar Properties of Emotion Verbs and Their Representation in WordNet
181 An Implementation of a System of Verb Relations in plWordNet 2.0
Maziarz, Marek 189 Approaching plWordNet 2.0
273 Automated Generation of Derivative Relations in the Wordnet Expansion Perspective
Meena, Akhlesh 79 A Study of the Sense Annotation Process: Man v/s Machine
Merabti, Tayeb 94 Combining Wordnet and Crosslingual multi-terminology health portal to access health information

author	page	paper
Miyake, Hirofumi	349	Extension of Phrases for Article Determination using WordNet Thesaurus
Mohanty, Panchanan	197	NOUNS IN ODIA: AN ONTOLOGICAL PERSPECTIVE
Mohanty, Sanghamitra	205	Ontology of Sanskrit Wordnet: Nouns and Verbs
Mok, Hazel	211	Using WordNet to predict numeral classifiers in Chinese and Japanese
Monachini, Monica	219	Mapping a corpus-induced ontology of action verbs on ItalWordNet
Moneglia, Massimo	219	Mapping a corpus-induced ontology of action verbs on ItalWordNet
Mukherjee, Subhabrata	10	Leveraging Sentiment to Compute Word Similarity
N		
Nagata, Ryo	349	Extension of Phrases for Article Determination using WordNet Thesaurus
Narayana, V.N	240	Kannada Verbs and their Automatic Sense Disambiguation
Niemi, Jyrki	227	Using a Bilingual Resource to Add Synonyms to a Wordnet: FinnWordNet and Wikipedia as an Example
Nishino, Tetsuro	35	Using WordNet to Handle the OOV Problem in English to Bangla Machine Translation
Nomorosa, Karen	248	Wordnet and SUMO for Sentiment Analysis
Novak, Jernej	113	sloWNet 3.0: development, extension and cleaning
O		
Oliver, Antoni	232	Building WordNets by machine translation of sense tagged corpora
Orav, Heili	254	Linking and Validating Nordic and Baltic Wordnets – A Multilingual Action in META-NORD
Oronoz Anchordoqui, Maite	260	Using WordNet into UKB in a Question Answering System for Basque
Ototake, Hokuto	349	Extension of Phrases for Article Determination using WordNet Thesaurus
P		
Paik, Kyonghee	64	A Survey of WordNets and their Licenses
Pala, Karel	50	Extending CzechWordNet Using a Bilingual Dictionary
Panunzi, Alessandro	219	Mapping a corpus-induced ontology of action verbs on ItalWordNet
Parameswarappa, S	240	Kannada Verbs and their Automatic Sense Disambiguation
Patel, Kirit	45	Introduction to Gujarati wordnet
Pease, Adam	248	Wordnet and SUMO for Sentiment Analysis
Pedersen, Bolette Sandford	254	Linking and Validating Nordic and Baltic Wordnets – A Multilingual Action in META-NORD
Perez De Viñaspre	260	Using WordNet into UKB in a Question Answering System for Basque
Garralda, Olatz		
Pianta, Emanuele	159	Encoding Commonsense Lexical Knowledge into WordNet

author	page	paper
Piasecki, Maciej	181	An Implementation of a System of Verb Relations in plWordNet 2.0
	189	Approaching plWordNet 2.0
	268	Automated Mapping of Polish Proper Names on plWordNet Hypernymy Structure
	273	Automated Generation of Derivative Relations in the Wordnet Expansion Perspective
Pokharel, Madhav	281	Mass noun classifiers in Nepali
Pääkkö, Paula	286	Finding a Location for a New Word in WordNet
Q		
Quattri, Francesca	294	Introducing WordNet in Interpreting Studies – Implications and Desiderata
R		
Rademaker, Alexandre	100	Revisiting a Brazilian WordNet
Rambousek, Adam	299	Low-cost ontology development
Ramocki, Radoslaw	305	Migrating Cornetto Lexicon to New XML Database Engine
Rigau, German	273	Automated Generation of Derivative Relations in the Wordnet Expansion Perspective
	72	Semantic Hand-Tagging of the SenSem Corpus Using Spanish WordNet Senses
	118	Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base
	382	Cross-lingual event-mining using wordnet as a shared knowledge interface
Roy, Arindam	312	A Proposed Nepali Synset Entry and Extraction tool
Russo, Irene	219	Mapping a corpus-induced ontology of action verbs on ItalWordNet
Rögnvaldsson, Eirikur	254	Linking and Validating Nordic and Baltic Wordnets – A Multilingual Action in META-NORD
S		
Sagot, Benoît	317	Automatic Extension of WOLF
Saharia, Navanath	149	An Extractive Approach of Text Summarization of Assamese using WordNet
Saikia, Utpal	324	A Novel Approach for Document Classification using Assamese WordNet
Sarkar, Sunita	155	ASSAMESE VOCABULARY AND ASSAMESE WORDNET BUILDING: AN ANALYSIS
Sarma, Shikhar K	312	A Proposed Nepali Synset Entry and Extraction tool
Sarmah, Jumi	324	A Novel Approach for Document Classification using Assamese WordNet
Schulam, Peter	324	A Novel Approach for Document Classification using Assamese WordNet
Shamsfard, Mehrnoush	330	Refining WordNet adjective dumbbells using intensity relations
Sharma, Utpal	344	Linking WordNet to DBpedia
Sheinman, Vera	149	An Extractive Approach of Text Summarization of Assamese using WordNet
	330	Refining WordNet adjective dumbbells using intensity relations

author	page	paper
Shyam Purkayastha, Bipul	312	A Proposed Nepali Synset Entry and Extraction tool
Soroa, Aitor	382	Cross-lingual event-mining using wordnet as a shared knowledge interface
Stevens, Keith	338	The C-Cat Wordnet Package: An Open Source Package for modifying and applying Wordnet
Szpakowicz, Stanisław	181	An Implementation of a System of Verb Relations in plWordNet 2.0
	189	Approaching plWordNet 2.0

T

Taheri, Aynaz	344	Linking WordNet to DBpedia
Takeuchi, Hiromi	349	Extension of Phrases for Article Determination using WordNet Thesaurus
Tesconi, Maurizio	23	WordNet Atlas: a web application for visualizing WordNet as a zoomable map
Tokunaga, Takenobu	330	Refining WordNet adjective dumbbells using intensity relations
Trouvilliez, Benoît	357	Linking specific and generalist knowledge
Tuveri, Franco	365	A Linguistic Approach to Feature Extraction Based on a Lexical Database of the Properties of Adjectives and Adverbs.

U

Uchimoto, Kiyotaka	56	Japanese SemCor: A Sense-tagged Corpus of Japanese
---------------------------	----	--

V

Vare, Kadri	167	Visual Study of Estonian Wordnet using Bipartite Graphs and Minimal Crossing algorithm
Veale, Tony	371	Teaching WordNet to Sing like an Angel and Cry like a Baby: Learning Affective Stereotypical Behaviors from the Web
Vincze, Veronika	377	Multiword Verbs in WordNets
Vivaldi, Jorge	390	Upgrading WordNet: a Terminological Point of View
Vohandu, Leo	167	Visual Study of Estonian Wordnet using Bipartite Graphs and Minimal Crossing algorithm
Vossen, Piek	382	Cross-lingual event-mining using wordnet as a shared knowledge interface

Y

Yamada, Setsuo	35	Using WordNet to Handle the OOV Problem in English to Bangla Machine Translation
-----------------------	----	--

Z

Zanotti, Cristian	390	Upgrading WordNet: a Terminological Point of View
Zapirain, Beñat	382	Cross-lingual event-mining using wordnet as a shared knowledge interface

author	page	paper
Zhou, Weitao	86	A Computer Aided Approach for Enriching WordNet with Semantic Definition
C		
Čapek, Tomáš	400	SENEQA – System for Quality Testing of Wordnet Data

Keyword Index

keyword paper

A

Abstracting and indexing	94	Combining Wordnet and Crosslingual multi-terminology health portal to access health information
accesibility	64	A Survey of WordNets and their Licenses
Action types	219	Mapping a corpus-induced ontology of action verbs on ItalWordNet
Adding synonyms	227	Using a Bilingual Resource to Add Synonyms to a Wordnet: FinnWordNet and Wikipedia as an Example
adjective scale	330	Refining WordNet adjective dumbbells using intensity relations
Adjectives	142	Restructuring Adjectives in WordNet with ClusterEditor
adjectives	330	Refining WordNet adjective dumbbells using intensity relations
affect	371	Teaching WordNet to Sing like an Angel and Cry like a Baby: Learning Affective Stereotypical Behaviors from the Web
animacy	197	NOUNS IN ODIA: AN ONTOLOGICAL PERSPECTIVE
Arabic	18	Toward Building a Large-Scale Arabic Sentiment Lexicon
ArchiWordNet	40	A new hierarchy of ArchiWordNet (AWN): building parts implementation with image
Argument Structure	240	Kannada Verbs and their Automatic Sense Disambiguation
article determination	349	Extension of Phrases for Article Determination using WordNet Thesaurus
Assamese	149	An Extractive Approach of Text Summarization of Assamese using WordNet
Assamese,	324	A Novel Approach for Document Classification using Assamese WordNet
	155	ASSAMESE VOCABULARY AND ASSAMESE WORDNET BUILDING: AN ANALYSIS
assisted lexicography	286	Finding a Location for a New Word in WordNet
Assistive use of WordNet	159	Encoding Commonsense Lexical Knowledge into WordNet
automatic wordnet development	113	sloWNet 3.0: development, extension and cleaning

B

Baltic	254	Linking and Validating Nordic and Baltic Wordnets – A Multilingual Action in META-NORD
Basque	260	Using WordNet into UKB in a Question Answering System for Basque
behavior	371	Teaching WordNet to Sing like an Angel and Cry like a Baby: Learning Affective Stereotypical Behaviors from the Web

keyword	paper
bilingual dictionary	50 Extending CzechWordNet Using a Bilingual Dictionary
bilingual lexica	317 Automatic Extension of WOLF
Bilingual resources	Using a Bilingual Resource to Add Synonyms to a Wordnet: FinnWordNet and Wikipedia as an Example 227
bipartite graphs	Visual Study of Estonian Wordnet using Bipartite Graphs and Minimal Crossing algorithm 167
Brazilian Portuguese	100 Revisiting a Brazilian WordNet
building	205 Ontology of Sanskrit Wordnet: Nouns and Verbs
BushBank	299 Low-cost ontology development
C	
case-suffix attachment	197 NOUNS IN ODIA: AN ONTOLOGICAL PERSPECTIVE
Catalan	232 Building WordNets by machine translation of sense tagged corpora
Cataloguing	Combining Wordnet and Crosslingual multi-terminology health portal to access health information 94
Chinese	211 Using WordNet to predict numeral classifiers in Chinese and Japanese
classification systems	173 Rethinking WordNet's Domains
classifier	281 Mass noun classifiers in Nepali
classifiers	211 Using WordNet to predict numeral classifiers in Chinese and Japanese
closed sets	Visual Study of Estonian Wordnet using Bipartite Graphs and Minimal Crossing algorithm 167
Clustering	142 Restructuring Adjectives in WordNet with ClusterEditor
cognitive analysis	281 Mass noun classifiers in Nepali
Commonsense Knowledge	159 Encoding Commonsense Lexical Knowledge into WordNet
CONJUNCT VERBS	30 Verbs in Sanskrit Wordnet
context	Teaching WordNet to Sing like an Angel and Cry like a Baby: Learning Affective Stereotypical Behaviors from the Web 371
Controlled Vocabulary	Combining Wordnet and Crosslingual multi-terminology health portal to access health information 94
Cornetto	305 Migrating Cornetto Lexicon to New XML Database Engine
Corpora	240 Kannada Verbs and their Automatic Sense Disambiguation
corpus	349 Extension of Phrases for Article Determination using WordNet Thesaurus
corpus-based approach	189 Approaching plWordNet 2.0
count noun classifier	281 Mass noun classifiers in Nepali
cross-lingual	382 Cross-lingual event-mining using wordnet as a shared knowledge interface

keyword	paper
cross-lingual semantic similarity	134 Computing Cross-Lingual Synonym Set Similarity by Using Princeton Annotated Gloss Corpus
Czech WordNet	50 Extending CzechWordNet Using a Bilingual Dictionary
D	
data sparseness	349 Extension of Phrases for Article Determination using WordNet Thesaurus
Database	94 Combining Wordnet and Crosslingual multi-terminology health portal to access health information
datamining of wikipedia	286 Finding a Location for a New Word in WordNet
DBpedia	344 Linking WordNet to DBpedia
derivational morphology	273 Automated Generation of Derivative Relations in the Wordnet Expansion Perspective
derivational relations	273 Automated Generation of Derivative Relations in the Wordnet Expansion Perspective
Dictionaries	94 Combining Wordnet and Crosslingual multi-terminology health portal to access health information
Dictionary	240 Kannada Verbs and their Automatic Sense Disambiguation
Document Classification	324 A Novel Approach for Document Classification using Assamese WordNet
E	
Editor	142 Restructuring Adjectives in WordNet with ClusterEditor
Europe	94 Combining Wordnet and Crosslingual multi-terminology health portal to access health information
evaluation	305 Migrating Cornetto Lexicon to New XML Database Engine
event mining	382 Cross-lingual event-mining using wordnet as a shared knowledge interface
Example-Based Machine Translation	35 Using WordNet to Handle the OOV Problem in English to Bangla Machine Translation
Expansion Approach	45 Introduction to Gujarati WordNet
expansion approach	312 A Proposed Nepali Synset Entry and Extraction Tool
EXPANSION	30 Verbs in Sanskrit Wordnet
extending wordnet	286 Finding a Location for a New Word in WordNet
extension	50 Extending CzechWordNet Using a Bilingual Dictionary
Extension of WordNet	159 Encoding Commonsense Lexical Knowledge into WordNet
extraction of lexico-semantic relations	268 Automated Mapping of Polish Proper Names on plWordNet Hypernymy Structure
Extractive	149 An Extractive Approach of Text Summarization of Assamese using WordNet

F

Feature Extraction	365	A Linguistic Approach to Feature Extraction Based on a Lexical Database of the Properties of Adjectives and Adverbs
Feature Norms	159	Encoding Commonsense Lexical Knowledge into WordNet
FinnWordNet	227	Using a Bilingual Resource to Add Synonyms to a Wordnet: FinnWordNet and Wikipedia as an Example

G

gender	281	Mass noun classifiers in Nepali
generalist resources	357	Linking specific and generalist knowledge
gradation	330	Refining WordNet adjective dumbbells using intensity relations
graph visualization	23	WordNet Atlas: a web application for visualizing WordNet as a zoomable map
GUI	142	Restructuring Adjectives in WordNet with ClusterEditor
Gujarati	45	Introduction to Gujarati WordNet

H

heuristic test	400	SENEQA – System for Quality Testing of Wordnet Data
Hungarian WordNet	377	Multiword Verbs in WordNets

I

illustrated WordNet	40	A new hierarchy of ArchiWordNet (AWN): building parts implementation with image
Images	219	Mapping a corpus-induced ontology of action verbs on ItalWordNet
IndoWordnet	45	Introduction to Gujarati WordNet
infographic	23	WordNet Atlas: a web application for visualizing WordNet as a zoomable map
Information Content	126	New WordNet-based semantic relatedness measurement: Using new information content metric and k-means clustering algorithm
Information Storage and Retrieval	94	Combining Wordnet and Crosslingual multi-terminology health portal to access health information
intensity relation	330	Refining WordNet adjective dumbbells using intensity relations
intensity scales	105	Scalar Properties of Emotion Verbs and Their Representation in WordNet
Internet	94	Combining Wordnet and Crosslingual multi-terminology health portal to access health information
interpreting studies	294	Introducing WordNet in Interpreting Studies – Implications and Desiderata

keyword	paper
---------	-------

J

- Japanese** 56 Japanese SemCor: A Sense-tagged Corpus of Japanese
Japanese learners of English 211 Using WordNet to predict numeral classifiers in Chinese and Japanese
349 Extension of Phrases for Article Determination using WordNet Thesaurus

K

- K-means** 126 New WordNet-based semantic relatedness measurement: Using new information content metric and k-means clustering algorithm
knowledge extraction 382 Cross-lingual event-mining using wordnet as a shared knowledge interface
knowledge integration 118 Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base
knowledge representation 100 Revisiting a Brazilian WordNet

L

- Languages for Specific Purposes** 390 Upgrading WordNet: a Terminological Point of View
lexical relations 181 An Implementation of a System of Verb Relations in plWordNet 2.0
189 Approaching plWordNet 2.0
lexical semantics 330 Refining WordNet adjective dumbbells using intensity relations
lexical units 189 Approaching plWordNet 2.0
LEXICALISATION 30 Verbs in Sanskrit Wordnet
license 64 A Survey of WordNets and their Licenses
LIGHT VERBS, 30 Verbs in Sanskrit Wordnet
linking 254 Linking and Validating Nordic and Baltic Wordnets – A Multilingual Action in META-NORD
Logic 100 Revisiting a Brazilian WordNet
low-cost 299 Low-cost ontology development
low-resource language 35 Using WordNet to Handle the OOV Problem in English to Bangla Machine Translation
LSA Ranking 86 A Computer Aided Approach for Enriching WordNet with Semantic Definition

M

- machine translation** 232 Building WordNets by machine translation of sense tagged corpora
mapping of proper names onto wordnet 268 Automated Mapping of Polish Proper Names on plWordNet Hypernymy Structure
mass noun classifier 281 Mass noun classifiers in Nepali

keyword	paper
matching	344 Linking WordNet to DBpedia
matching of lexicalized concepts	134 Computing Cross–Lingual Synonym Set Similarity by Using Princeton Annotated Gloss Corpus
measure of semantic relatedness	268 Automated Mapping of Polish Proper Names on plWordNet Hypernymy Structure
methodology	299 Low–cost ontology development
minimal crossing algorithm	167 Visual Study of Estonian Wordnet using Bipartite Graphs and Minimal Crossing algorithm
morphosemantic relations	273 Automated Generation of Derivative Relations in the Wordnet Expansion Perspective
multi-word units	294 Introducing WordNet in Interpreting Studies – Implications and Desiderata
Multilingual Central Repository	118 Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base
Multilingual wordnet	390 Upgrading WordNet: a Terminological Point of View
multilinguality	377 Multiword Verbs in WordNets
multiword expressions	377 Multiword Verbs in WordNets

N

Natural Language Processing	365 A Linguistic Approach to Feature Extraction Based on a Lexical Database of the Properties of Adjectives and Adverbs
natural language processing	211 Using WordNet to predicting Numeral Classifiers in Chinese and Japanese
Nordic	254 Linking and Validating Nordic and Baltic Wordnets – A Multilingual Action in META–NORD
noun class	281 Mass noun classifiers in Nepali
numeral classifier	281 Mass noun classifiers in Nepali

O

Odia	197 NOUNS IN ODIA: AN ONTOLOGICAL PERSPECTIVE
	248 Wordnet and SUMO for Sentiment Analysis
ontology	299 Low–cost ontology development
	382 Cross–lingual event–mining using wordnet as a shared knowledge interface
Ontology	197 NOUNS IN ODIA: AN ONTOLOGICAL PERSPECTIVE
	219 Mapping a corpus–induced ontology of action verbs on ItalWordNet
Open Source	338 The C–Cat Wordnet Package: An Open Source Package for modifying and applying Wordnet
OpinionFinder	248 Wordnet and SUMO for Sentiment Analysis

keyword	paper
Out-Of-Vocabulary words	35 Using WordNet to Handle the OOV Problem in English to Bangla Machine Translation
P	
pattern-based extraction	268 Automated Mapping of Polish Proper Names on plWordNet Hypernymy Structure
PERI-PHRASES	30 Verbs in Sanskrit Wordnet
plWordNet	181 An Implementation of a System of Verb Relations in plWordNet 2.0 189 Approaching plWordNet 2.0
	181 An Implementation of a System of Verb Relations in plWordNet 2.0 189 Approaching plWordNet 2.0
Polish	273 Automated Generation of Derivative Relations in the Wordnet Expansion Perspective
	268 Automated Mapping of Polish Proper Names on plWordNet Hypernymy Structure
Polish wordnet	273 Automated Generation of Derivative Relations in the Wordnet Expansion Perspective
possessive classifier	281 Mass noun classifiers in Nepali
Princeton Annotated Gloss Corpus	134 Computing Cross-Lingual Synonym Set Similarity by Using Princeton Annotated Gloss Corpus
Princeton WordNet	377 Multiword Verbs in WordNets
Psycholinguistic foundations of WordNet	159 Encoding Commonsense Lexical Knowledge into WordNet
Q	
QA	260 Using WordNet into UKB in a Question Answering System for Basque
Quality assessment	72 Semantic Hand-Tagging of the SenSem Corpus Using Spanish WordNet Senses
quality testing	400 SENEQA – System for Quality Testing of Wordnet Data
R	
relation borrowing	312 A Proposed Nepali Synset Entry and Extraction Tool
resources	18 Toward Building a Large-Scale Arabic Sentiment Lexicon
S	
sales catalogs	357 Linking specific and generalist knowledge
Sanskrit	205 Ontology of Sanskrit Wordnet: Nouns and Verbs

keyword	paper
Semantic annotation	72 Semantic Hand-Tagging of the SenSem Corpus Using Spanish WordNet Senses
semantic annotation of corpora	113 sloWNNet 3.0: development, extension and cleaning
Semantic Definition	86 A Computer Aided Approach for Enriching WordNet with Semantic Definition
semantic hierarchies	211 Using WordNet to predict numeral classifiers in Chinese and Japanese
semantic interoperability	382 Cross-lingual event-mining using wordnet as a shared knowledge interface
Semantic pattern	240 Kannada Verbs and their Automatic Sense Disambiguation
semantic relations	167 Visual Study of Estonian Wordnet using Bipartite Graphs and Minimal Crossing algorithm
semantic similarity	260 Using WordNet into UKB in a Question Answering System for Basque
Semantic Similarity	126 New WordNet-based semantic relatedness measurement: Using new information content metric and k-means clustering algorithm
SemCor	56 Japanese SemCor: A Sense-tagged Corpus of Japanese
Sense Annotation Process	79 A Study of the Sense Annotation Process: Man v/s Machine
sense tagged corpus	232 Building WordNets by machine translation of sense tagged corpora
Sensebank	56 Japanese SemCor: A Sense-tagged Corpus of Japanese
SenSem	72 Semantic Hand-Tagging of the SenSem Corpus Using Spanish WordNet Senses
sentiment	18 Toward Building a Large-Scale Arabic Sentiment Lexicon
Sentiment	10 Leveraging Sentiment to Compute Word Similarity
Sentiment Analysis	365 A Linguistic Approach to Feature Extraction Based on a Lexical Database of the Properties of Adjectives and Adverbs
sentiment analysis	248 Wordnet and SUMO for Sentiment Analysis
Similarity metrics	10 Leveraging Sentiment to Compute Word Similarity
Snese Projection	56 Japanese SemCor: A Sense-tagged Corpus of Japanese
Spanish	232 Building WordNets by machine translation of sense tagged corpora
Spanish WordNet	72 Semantic Hand-Tagging of the SenSem Corpus Using Spanish WordNet Senses
specific knowledge	357 Linking specific and generalist knowledge
Statistics of Error analysis	79 A Study of the Sense Annotation Process: Man v/s Machine
Subject Headings	94 Combining Wordnet and Crosslingual multi-terminology health portal to access health information
subject-verb agreement	197 NOUNS IN ODIA: AN ONTOLOGICAL PERSPECTIVE
subjectivity	18 Toward Building a Large-Scale Arabic Sentiment Lexicon
SUFFIXATION,	30 Verbs in Sanskrit Wordnet
SUMO	248 Wordnet and SUMO for Sentiment Analysis
supervised learning	273 Automated Generation of Derivative Relations in the Wordnet Expansion Perspective
synset	312 A Proposed Nepali Synset Entry and Extraction Tool

keyword	paper
----------------	--------------

T

Taxonomy resources	173 Rethinking WordNet's Domains
Terminology	390 Upgrading WordNet: a Terminological Point of View
Terminology as subject.	94 Combining Wordnet and Crosslingual multi–terminology health portal to access health information
Text Summarization	149 An Extractive Approach of Text Summarization of Assamese using WordNet
Thematic role	240 Kannada Verbs and their Automatic Sense Disambiguation
thesaurus	40 A new hierarchy of ArchiWordNet (AWN): building parts implementation with image
translation	50 Extending CzechWordNet Using a Bilingual Dictionary

U

UKB	260 Using WordNet into UKB in a Question Answering System for Basque
usability	64 A Survey of WordNets and their Licenses

V

validation	254 Linking and Validating Nordic and Baltic Wordnets – A Multilingual Action in META–NORD
Verb Sense Disambiguation	240 Kannada Verbs and their Automatic Sense Disambiguation
verbal classifier	281 Mass noun classifiers in Nepali
Verbalizer	240 Kannada Verbs and their Automatic Sense Disambiguation
verbs	181 An Implementation of a System of Verb Relations in piWordNet 2.0
verbs of emotion	105 Scalar Properties of Emotion Verbs and Their Representation in WordNet
Vocabulary	155 ASSAMESE VOCABULARY AND ASSAMESE WORDNET BUILDING: AN ANALYSIS

W

web application	23 WordNet Atlas: a web application for visualizing WordNet as a zoomable map
	227 Using a Bilingual Resource to Add Synonyms to a Wordnet: FinnWordNet and Wikipedia as an Example
Wikipedia	268 Automated Mapping of Polish Proper Names on piWordNet Hypernymy Structure
word sense disambiguation	317 Automatic Extension of WOLF
Word Sense Disambiguation	79 A Study of the Sense Annotation Process: Man v/s Machine

keyword	paper
	338 The C-Cat Wordnet Package: An Open Source Package for modifying and applying Wordnet
	365 A Linguistic Approach to Feature Extraction Based on a Lexical Database of the Properties of Adjectives and Adverbs
	167 Visual Study of Estonian Wordnet using Bipartite Graphs and Minimal Crossing algorithm
	181 An Implementation of a System of Verb Relations in plWordNet 2.0
	189 Approaching plWordNet 2.0
wordnet	
	205 Ontology of Sanskrit Wordnet: Nouns and Verbs
	305 Migrating Cornetto Lexicon to New XML Database Engine
	330 Refining WordNet adjective dumbbells using intensity relations
	382 Cross-lingual event-mining using wordnet as a shared knowledge interface
	400 SENEQA – System for Quality Testing of Wordnet Data
	10 Leveraging Sentiment to Compute Word Similarity
	23 WordNet Atlas: a web application for visualizing WordNet as a zoomable map
	35 Using WordNet to Handle the OOV Problem in English to Bangla Machine Translation
	45 Introduction to Gujarati WordNet
	100 Revisiting a Brazilian WordNet
WordNet	
	126 New WordNet-based semantic relatedness measurement: Using new information content metric and k-means clustering algorithm
	142 Restructuring Adjectives in WordNet with ClusterEditor
	248 Wordnet and SUMO for Sentiment Analysis
	260 Using WordNet into UKB in a Question Answering System for Basque
	312 A Proposed Nepali Synset Entry and Extraction Tool
	324 A Novel Approach for Document Classification using Assamese WordNet
WordNet 3.0	
wordnet applications	
wordnet browser, editor and visualization tool	
wordnet construction	
WordNet domain ontology	
Wordnet domains	
Wordnet enrichment	

keyword	paper
wordnet expansion	Automated Mapping of Polish Proper Names on plWordNet Hypernymy Structure 268
	Automated Generation of Derivative Relations in the Wordnet Expansion Perspective 273
wordnet extension	Automatic Extension of WOLF 317
Wordnet library	The C-Cat Wordnet Package: An Open Source Package for modifying and applying Wordnet 338
wordnet of proper names	Automated Mapping of Polish Proper Names on plWordNet Hypernymy Structure 268
WordNet representation	Scalar Properties of Emotion Verbs and Their Representation in WordNet 105
Wordnet similarity	The C-Cat Wordnet Package: An Open Source Package for modifying and applying Wordnet 338
wordnet structure	Multiword Verbs in WordNets 377
WordNet thesaurus	Extension of Phrases for Article Determination using WordNet Thesaurus 349
Wordnet,	ASSAMESE VOCABULARY AND ASSAMESE WORDNET BUILDING: AN ANALYSIS 155
wordnets	Linking and Validating Nordic and Baltic Wordnets – A Multilingual Action in META-NORD 254
	Introducing WordNet in Interpreting Studies – Implications and Desiderata 294
WorNet	Mapping a corpus-induced ontology of action verbs on ItalWordNet 219
WornNet Enriching	A Computer Aided Approach for Enriching WordNet with Semantic Definition 86
X	
XML database	Migrating Cornetto Lexicon to New XML Database Engine 305
Z	
zooming interface	WordNet Atlas: a web application for visualizing WordNet as a zoomable map 23

GWC 2012

Proceedings of the 6th International Global Wordnet Conference

Editors: Christiane Fellbaum, Piek Vossen

Printed and published by Tribun EU s. r. o.

Cejl 32, 602 00 Brno, Czech Republic

First edition at Tribun EU

Brno 2012

ISBN 978-80-263-0244-5

www.librix.eu