Automatic Acquisition of Controlled Vocabularies from Wikipedia using Wikilinks, Word Ranking, and a Dependency Parser

No Author Given

No Institute Given

Abstract. Controlled vocabularies are important resources used in several tasks such as machine translation, text summarization, and text analysis. However, the development of such resources is expensive and time-consuming. On the other hand, the Wikipedia, a free collaborative encyclopedia, contains plenty of semi-structured information that can be used by an automatic process to create new resources. This paper proposes a method to extract semantic information from the Wikipedia in the form of a controlled vocabulary. The method combines keywords obtained for a specific Wikipedia article with three different strategies: using Wikipedia annotations called wikilinks, a ranking measure to obtain keywords from text, and a dependency parser. To evaluate the model, we performed an analysis in terms of coverage and performance of the acquired vocabulary using WordNet as a gold standard.

1 Introduction

Wikipedia is a free-access encyclopedia, written by all people over the world. Wikipedia is targeted to a general public that can collaborate by improving its content through a web interface. Users of Wikipedia edit the content by adding new text that includes annotations with information such as links between articles and categories. These annotations turn a Wikipedia article into a complex semantic network of inter-related terms.

Wikipedia's articles contain different sort of metadata annotations: content tables, categories, inter-language links, and Wikipedia markup annotations in the text such as links to internal pages, called freelinks. That information has been used previously for knowledge discovery using the Wikipedia. Some examples include mining information from content tables or Wiki infoboxes [8], taxonomy deriving from Wiki categories [18], and topic hierarchies from Wiki categories [4].

In this study, we analyze the use of a Wikipedia to obtain a set of keywords given a specific article. We explore three different methods to acquire semantic related words. Such set of keywords can be seen as an entry in a semantic dictionary or a controlled vocabulary, where the title of the Wikipedia article is the concept and the terms related to the concept are the keywords. We argue that such set of semantically related keywords can be useful in many scenarios,

such as text classification where they can be used as the features for a classifier to reduce the dimensionality. We propose to acquire a set of words is, in fact, a controlled vocabulary, where the name of the article is the main term and text and annotations on the article contain the edges to other nodes. For example, given a set of terms such as *Christianity*, *Bible*, *Computer*, and *Operating system*, it is possible to acquire a set of related terms for each of those articles. This controlled vocabulary can be used for different tasks like text classification, where each of the terms is related to one of the following categories: *Religion* or *Computers*. Then, the categorized vocabulary can be used as feature reduction method to classify the documents, reducing the time an usual method need to analyze a document since the possible set of features is being reduced.

In this study, we focus on using three different methods to obtain keywords from a specific Wikipedia article. The first one is using annotations to other articles in a Wikipedia article called wikilinks. The second one is extracting keywords from the text taking into account the frequency of the term in the article. The thirds one is also using the text to find syntactic relations between a particular term in the sentences. Finally, we propose a method to combine them all, improving the precision of the quality set of keywords.

We experimented with the information provided by a set of specific articles and compared the results with the well know resource called WordNet. WordNet organizes terms as a group of synonyms called synset, similar to a thesaurus. We used WordNet to measure the quality of the acquired set of words for each article by comparing it with the terms found on the WordNet synset.

The rest of this article is explained as follows. Section 2 provides information about the related work in information extraction using the Wikipedia. Section 3 contains an explanation of the three methods used to extract keywords. Empirical results are presented and discussed in section 4, while section 5 presents the conclusions.

2 Related works

Recently, Wikipedia has been used as a source of information. To give some examples, [10] describes a method to obtain semantic information for word sense disambiguation. [9] presents a system that trains himself with data from Wikipedia to perform text annotation and add semantic information to plan texts. [2] extracts taxonomies from Wikipedia. Other works have proved that information such as hyperlink relations between a set of documents [1] can be used to obtain a set of interrelated terms. Another kind of studies have used hyperlinks in the Wikipedia provides more information than in a common website [11].

Dependency parsing has proven to be a method to acquire semantic patterns because they are capable of representing relations between elements of a sentence. For example, a dependency parser has been used to infer rules for automatic question answering [7] and paraphrase identification [15]. Generally, these systems have been focused on relation extraction, the identification of particular

relations between items in the text. The relations used and the method to acquire them vary across studies. For example, [17] used an [s,v,o] (subject, verb, direct object) tuples as patterns while [14] allow any analysis to be a possible pattern. Machine learning algorithms are also used to learn and identify instances of relations such as iterative semi-supervised algorithms [3] where usually the algorithm is provided with an initial set of patterns and attempts to increase the number of patterns and relations between items. This work, however, is more related to the work of [14, 13], in the sense it uses different sources to rank terms obtained by a dependency parser. [14] used tf-idf method to rank patterns according to where they tended to occur in documents which were known to contain information of interest while [13] proposed a ranking term measure to give a value to terms obtained by patterns obtained from the result of a dependency parser.

3 Acquiring a Controlled Vocabulary from Wikipedia

This section describes the proposed process to acquire a controlled vocabulary from the Wikipedia. First, we describe the general process to acquire a set of keywords related to a specific article using three approaches and then mixing them. Then, we explain each one of the three approaches: wikilinks, word ranking, and dependency parsing.

We define a controlled vocabulary as a list of concepts, also called entries, where each of them has selected list of words semantically related to such concept. A controlled vocabulary provides a method to organize knowledge. It is used in subject indexing schemes, subject headings, thesauri, taxonomies and other forms of knowledge organization systems. A controlled vocabulary specifies the use of terms and phrases related to a concept, similar to a dictionary, where each of the entries has a set related terms. In this case, we are interested in acquiring only the terms related to a concept. For example, given the concept 'cat', the task is to acquire a set of words that are related such as 'animal', 'eat', or 'kitty'.

We propose to acquire a controlled vocabulary for specific terms using articles from the Wikipedia. Figure 1 depicts the extraction process. Having an article from Wikipedia, two different extraction processes are performed: extraction of terms using wikilinks (1) and text extraction (2). Extraction of terms using wikilinks is explained in detail in Section 3.1, while text extraction consists in extracting the text of the article. Extraction of terms using term frequency (3) and a dependency parser (4) is performed on the extracted text. Both processes are explained in Section 3.2 and Section 3.4 respectively. Finally, the three set of keywords is combined to finally obtain one set of keywords (5). The combination process is explained in Section 3.4.

3.1 Extraction of terms using Wikilinks

The structure of Wikipedia can be exploited in several ways. In particular, we are interested in using the so-called *wikilinks*. A wikilink is a reference to different

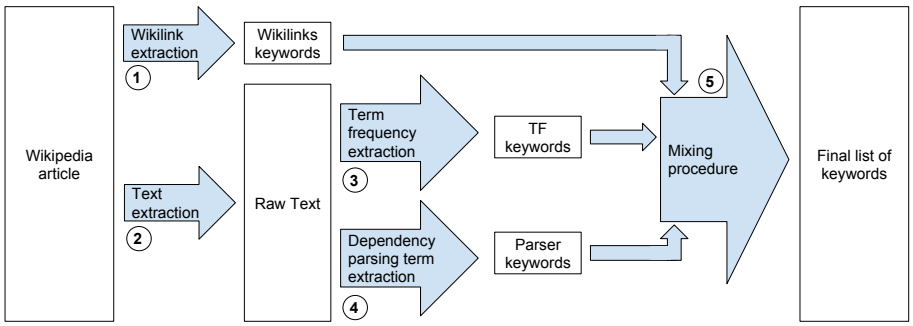

Fig. 1. Pipeline of the proposed method to extract the controlled vocabulary.

Wikipedia article. For example, the following wikitext has been extracted from the article *Operating system*:

> Examples of popular desktop operating systems include `[[Apple Inc.|Apple]] [[OS X]], [[Linux]]` and its variants, and `[[Microsoft Windows]]`. So-called `[[mobile operating system]]`s include `[[Android (operating system)|Android]]` and `[[iOS]]`.

It can be seen in the above example that it is possible to extract valuable keywords such as *Linux*, *Microsoft Windows*, and *Android*, represented by wikilinks. Furthermore, this information can be easily extracted since wikilinks maintain a regular pattern in the source of the article. One of the questions addressed in this article is how much of this information provided by all the wikilinks of a specific article is closely related to the semantic meaning of the article.

3.2 Term acquisition using frequency ranking

Frequency word ranking has been used to distinguish important terms in texts under the assumption that the importance of a term that occurs in a document is proportional to its term frequency. In other words, a term that frequently appears in the text is supposed to be important. The frequency of a term t in a document a, in this case, a Wikipedia article, is calculated as follows:

$$f(t,a) = f_{t,a} = count(t,a),$$

where the $count(t,a)$ is number of times the term t occurs in the document a.

However, in cases where the length of documents vary adjustments and normalizations are often made. We use the following measures to evaluate the usefulness of a term t_i:

– Term Frequency score (TF) :

$$\frac{f_{t,a}}{\max\{f_{t',a} : t' \in a\}},$$

where $f_{t,a}$ is the number of times the term t appears on the article a.
– Scaled Term Frequency score (STF) :

$$(1-k) + k\frac{f_{t,a}}{\max\{f_{t',a} : t' \in a\}},$$

where $f_{t,a}$ is the number of times the term t appears on the article a and $k \in R^+, 0 < k \leq 1$.
– Log-scaled Term Frequency score (LTF):

$$\frac{\log(f_{t,a})}{\log(\max\{f_{t',a} : t' \in a\})}$$

The term frequency score is the normalized count of the terms, where the value of term that occurs the most is 1 and a non important term in the document is close to 0. The scaled term frequency score provides a method to change the minimum weight of the frequency score using the k parameter. In other words, the STF score transforms any given value of the TF score from the range $[0, 1]$ to $[k, 1]$. The scale is k equal to 0.5. Finally, the log-scaled term frequency score allows to normalize a given value penalizing low occurrences. It worth mentioning that the normalized and scaled scores only work when combining the scores with other functions or to compare ranked works in different documents. If the purpose is to extract words in a document, the transformations will not help since they are all monotonic functions.

3.3 Dependency parsing

A dependency parser is a computational tool that takes a sentence as input and returns a tree, called a parse tree, that contains connections between words according to their relationships. In the tree, each node represents a word, child nodes are words that are dependent on the parent, and edges are labeled according to the relationship between the parent an the child. The set of relationships depends on the grammar. For example, The Universal Dependencies (UD) project defines a set of relations that by developing cross-linguistically treebank annotation for several languages [12]. Under such grammar, there is a relation called *nsubj* between two words $w1 \rightarrow w2$ that represents the case when $w2$ is a noun that is the syntactic subject of the clause. As an example, in the phrase 'Colombia won the tournament', the word Colombia' is related to the verb 'won' through the relation *nsubj*. Figure 2 shows the result of the parse tree of such sentence.

We propose the use of three relations to consider possible terms as candidates:

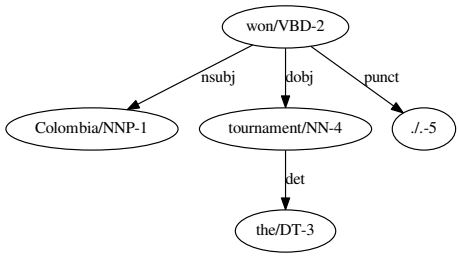

Fig. 2. An example of the dependency tree for the sentence 'Colombia won the tournament'.

- Any word that is the dominant of a nominal subject (nsubj) and is a verb and is related to the target concept. In Figure 2 the word 'won' would be selected. The objective of this relation is to find verbs.
- Any word that is a noun, has a dominant relation of any type with another node m, and m is the target concept. In Figure 2 the word 'tournament' would be selected. The objective of this relation is to find nouns.
- Any word that is a noun and also has dominant relation of any type with any node. The objective of this relation is also to find nouns.

3.4 Mixing sets of terms

The last part of the proposed method consists of combining the three previous methods to obtain keywords into one. We propose such combination by transforming each of the three previous methods into a normalized score and using a set of mixing coefficients, one for each of the scores. In the case of keywords extracted using the wikilink approachx, we propose the usage of an indicator function $WL(t, a)$ that returns 1 if the article a contains a wikilink with the term t and 0 if it does not. For the term frequency score for a particular document $WF(t, a)$, we propose the use one of the scores defined in Subsection 3.2. For the third one, the dependency parser usage, we also propose to use an indicator function $DP(t, a)$ that returns 1 if any of the relations listed in Section regarding the term t is found in the article a and 0 if it does not. Finally, a set of mixing coefficients α, β, and γ are used to control the importance of each score. The formal mathematical definition is stated as follows:

$$rank(t, a) = \alpha WL(t, a) + \beta WF(t, a) + \gamma DP(t, a),$$

where $\alpha \in R^+$, $\beta \in R^+$, and $\gamma \in R^+$, $\alpha + \beta + \gamma = 1$, are the mixing coefficients, $WL(t, a)$ is wikilinks indicator score, $WF(t, a)$ is a word frequency score, and $DP(t, a)$ is the dependency score.

3.5 Evaluation measures

In order to test the performance of our method, we propose the use of *precision* and *recall* measures. More specifically, given this set of target terms for a specific article $S = \{s_1, s_2, ..., s_K\}$, the precision of a set of terms $T_a = \{t_1, t_2, ..., t_N\}$ obtained for an article a related to S is calculated as:

$$precision(T_a) = \frac{count(\{t' \in S\})}{N},$$

the recall as:

$$recall(T_a) = \frac{count(\{t' \in S\})}{K},$$

and F1 score as:

$$F1score(T_a) = 2 * \frac{precision * recall}{precision + recall}$$

4 Experimentation and results

We used the Wikipedia dump, which is a file containing all the information stored in the Wikipedia in a specific time. To test our method, we used the also publicly available and well-known resource WordNet. WordNet organizes terms as a group of synonyms called synset. Each entry is called a synset and can be acquired using its surface word similar to a thesaurus. We used WordNet to measure the quality of the acquired set of words. For each selected article, we searched for all the possible terms that belong to the synset and that also are in the article. For example, given the article 'Sport', we included the term 'champion' if it is found in the Wikipedia article and also in the WordNet synset for 'sport'.

There are different measures to obtain synsets in the Wikipedia. We selected its the top terms using Wu-Palmer Similarity (WUP) [16] and Leacock-Chodorow Similarity (LCS) [6] with a threshold of 0.7. Finally, we used precision, recall, and F1 scores to evaluate the performance of the different strategies, as described in Subsection 3.5.

For experimentation, we selected manually 18 articles. Each article was preprocessed to extract all the information described previously: the set of wikilinks and text from a set of preselected articles. We used the publicly available library *wikitools* to extract the text and wikilinks from a set of articles[1]. With the text, we removed stopwords, punctuation, and transformed the text to lower case. Then, we selected two set keywords for each article using term frequency and a dependency parser. We used the Stanford Dependency Parser [5] as dependency parser in our experimentation. Finally, we measured how many of these terms are in the selected subset of terms from its WordNet synset.

For the experimentation, we selected a set of 18 articles. The full list can be seen in Table 1, along with the results. The first column shows the name

[1] The library can be found the in the git repository: https://github.com/rdorado79/wikitools

Concept	Count	Wikilinks			TF-score			Parser		
		Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Sport	177	0.083	0.107	0.094	0.155	**0.232**	**0.186**	**0.313**	0.028	0.052
Religion	128	0.076	**0.320**	0.123	0.120	0.180	**0.144**	**0.185**	0.094	0.124
Baseball	29	0.016	**0.276**	0.031	**0.093**	0.138	**0.111**	0.051	0.103	0.068
Soccer	9	0.025	**0.667**	0.049	0.308	0.444	**0.364**	**0.400**	0.222	0.286
CPU	34	0.009	0.059	0.016	**0.078**	**0.118**	**0.094**	0.068	0.088	0.077
Horse	95	0.070	**0.305**	0.114	0.056	0.084	0.068	**0.153**	0.095	**0.117**
Country	46	0.136	0.174	0.152	**0.159**	**0.239**	**0.191**	0.125	0.022	0.037
Computer	114	0.045	**0.158**	0.071	0.094	0.140	**0.112**	**0.109**	0.053	0.071
God	108	0.074	**0.278**	0.117	0.117	0.176	**0.141**	**0.189**	0.093	0.124
Dog	141	0.044	0.099	0.061	0.076	**0.113**	0.091	**0.164**	0.071	**0.099**
Planet	66	0.064	**0.379**	0.109	0.172	0.258	**0.206**	**0.200**	0.167	0.182
Jehovah	31	0.032	**0.226**	0.055	**0.065**	0.097	**0.078**	0.036	0.032	0.034
Music	269	0.093	**0.294**	0.141	0.141	0.212	**0.170**	**0.261**	0.112	0.156
Canada	35	0.013	**0.257**	0.024	**0.038**	0.057	0.046	0.036	0.086	**0.050**
Animal	171	**0.132**	**0.234**	**0.169**	0.070	0.105	0.084	0.000	0.000	NaN
Colombia	29	0.016	**0.448**	0.030	0.070	0.103	0.083	**0.111**	0.207	**0.145**
Cat	106	0.051	**0.198**	**0.082**	0.038	0.057	0.045	**0.060**	0.028	0.038
Japan	61	0.016	**0.180**	0.029	0.055	0.082	0.066	**0.059**	0.082	**0.068**
Average	91.61	0.055	**0.259**	0.081	0.106	0.157	**0.127**	**0.140**	0.088	0.096

Table 1. Results of the experimentation in counts

of the article. The second column shows the number of targeted words for the specific article. Columns 3 to 5 show the results for wikilinks, 6 to 8 for term frequency, and columns 9 to 11 for the dependency parser. It can be concluded that Wikilink term extraction gives the best recall between the three methods. This suggests that the keywords obtained using the wikilinks contain similar semantic information to Wordnet. In terms of precision, the keywords obtained by the dependency parser strategy seem to correlate WordNet slightly better. However, the best F1-score on average is obtained by the keywords obtained using the term frequency measure.

Despite the relatively low values, a thorough analysis of the extracted word shows encouraging results. Table 2 depicts examples of the extracted words for each one of the three methods. It can be seen the nature of the words extracted by each method. Terms extracted from wikilinks tend to be concepts that complement the Wikipedia article since wikilinks are links to other articles. Examples of these are 'tie-breaking' for the article 'Sport', or 'equus ferus' for 'Horse'. Terms extracted using the frequency are generally those that are used to explain important concepts in the document and appear frequently. Examples of these are 'faith' in the article 'Religion' and 'territory' in 'Country'. Finally, terms extracted with the dependency parser are generally functional words related to the concept. For example, the word 'movement' for 'Horse' or the word 'affect' for religion. It worths mentioning that the patterns used were very simple and are susceptible to improvement in many ways. This is left for future work.

Concept	Wikilinks	TF score	Dependency
Sport	competitive, physical activity, game, entertainment, tie-breaking methods, tournament, champion, playoffs	participants, physical, participation, used, including, competition, increase, football, international, games	charter, technology, participation, venue, popularity, definition, set, noun, running, spectator, football
Religion	cultural system, behaviors, world view, sacred texts, societal organisation, human, existence, divine	world, one, people, law, study, faith, belief, beliefs, include, sacred, god, defined, science, countries	view, essence, king, edition, violence, absence, excess, influence, part, role, affects, morality, deals, science
CPU	electronic circuit, computer, instruction, computer program, input/output, control unit, main memory, design	instruction, cpus, clock, instructions, program, memory, data, performance, operation, processors, execution	dissipation, cycle, context, perform, hardware, performance, designs, design, parallelism, operation, execute
Horse	extant, subspecies, equus ferus, mammal, equidae, evolved, eohippus, domesticate, anatomy, colors	breeds, developed, wild, domesticated, bones, four, animals, well, human, age, world	tibia, blood, movement, appears, style, control, rider, breed, pedigree, domestication, remains, zebra, animal
Country	political geography, sovereign state, political division, people, league of nations, united nations, states	country, sovereign, state, word, political, independent, associated, territory, english, govpubs, gov, derived	version, coal, part, sense, name, resident, sovereign state, region

Table 2. Random examples of the words extracted for some articles

The last part of the method to combine all the previous acquired set of words into one unique set. We combined all the three previous strategies by the method proposed in Section 3.5. Testing with different values for α, β, and γ, we found that appropriate values for the model are a high β and a γ value slightly higher than α. This is possible due that the TF score is lower than the other two and tend to be ruled out if it is not properly scaled and weighted. Also, a higher γ allows preference from the dependency parser, since they have a high precision.

Table 3 shows the results with $\alpha = 0.08$, $\beta = 0.8$, and $\gamma = 0.12$ for the mixing model. We also set $k = 0.8$ for the k parameter of the scaled TF score. It can be seen in the table that we successfully combined all the three strategies, wikilinks, term frequency, and the dependency parser, into one method obtaining an increment in the F1-score. It is also difficult to say which of the TF score used is better than other. Numerically, the scaled TF score outperforms the other two but log scaled TF score gives close values to the best found.

Concept	TF score			Scaled TF score			Log TF score		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Sport	0.121	0.215	0.155	0.137	0.243	0.175	**0.153**	**0.271**	**0.196**
Religion	**0.158**	**0.227**	**0.186**	**0.158**	**0.227**	**0.186**	0.119	0.180	0.143
Baseball	0.059	0.103	0.075	**0.136**	**0.207**	**0.164**	0.133	**0.207**	0.162
Soccer	0.182	0.222	0.200	**0.364**	**0.444**	**0.400**	**0.364**	**0.444**	**0.400**
CPU	0.078	0.118	0.094	**0.096**	**0.147**	**0.116**	0.077	0.118	0.093
Horse	**0.117**	**0.179**	**0.142**	**0.117**	**0.179**	**0.142**	0.106	0.168	0.130
Country	0.169	0.283	0.211	**0.182**	**0.304**	**0.228**	0.141	0.239	0.177
Computer	0.081	0.123	0.098	0.089	0.140	0.109	**0.108**	**0.175**	**0.134**
God	**0.170**	**0.241**	**0.199**	**0.170**	**0.241**	**0.199**	0.168	**0.241**	0.198
Dog	0.086	0.142	0.107	**0.091**	**0.149**	**0.113**	0.082	0.135	0.102
Planet	**0.212**	**0.273**	**0.238**	0.202	**0.273**	0.232	0.196	**0.273**	0.228
Jehovah	0.053	0.065	0.058	0.053	0.065	0.058	**0.077**	**0.097**	**0.086**
Music	**0.170**	0.227	0.194	**0.170**	0.227	0.194	0.169	**0.230**	**0.195**
Canada	0.063	**0.086**	0.072	**0.068**	**0.086**	**0.076**	0.037	0.057	0.045
Animal	**0.116**	**0.187**	**0.143**	0.083	0.135	0.103	0.080	0.129	0.098
Colombia	0.167	**0.207**	0.185	**0.188**	**0.207**	**0.197**	0.150	**0.207**	0.174
Cat	**0.045**	**0.075**	**0.057**	**0.045**	**0.075**	**0.057**	0.033	0.057	0.042
Japan	**0.074**	**0.115**	**0.090**	**0.074**	**0.115**	**0.090**	0.072	**0.115**	0.089
Average	0.118	0.171	0.139	**0.135**	**0.192**	**0.158**	0.126	0.185	0.149

Table 3. Results of mixing method with $\alpha = 0.08$, $\beta = 0.8$, and $\gamma = 0.12$ and $k = 0.8$ for the scaled TF score.

5 Conclusions

We presented a method to obtain a set of keywords, called a controlled vocabulary, from Wikipedia articles. The method consists in the combination of other three different methods to obtain keywords: wikilinks, term frequency, and a dependency parser. The presented method successfully combine all the three strategies, wikilinks, term frequency, and the dependency parser, into one achieving an increment in the F1-score.

In relation with the words extracted from each one of the methods, we can say that they differ substantially. As discussed in the results section, terms extracted from wikilinks tend to be additional concepts that complement the article. Terms extracted using the frequency are keywords that are used to explain important concepts in the document. Finally, terms extracted with the dependency parser are generally functional words related to the concept.

With respect to the dependency parser, we used three different relations: direct object of the term (dobj) to obtain verbs and two based on nominal subjects (nsubj) to obtain nouns. The method showed to be very promising in terms of precision and it worth improve the patterns in the future.

We expect to extend this work in many ways. First, we expect to perform a deeper analysis that involves a much higher number of articles. Second, we would like to explore the idea of acquiring automatically grammatical rules for dependency parsing based term acquisition. Third, we would want to perform a

deeper analysis on the extraction of keywords using wikilinks. We used a binary function that does not allow to rank and compare words using such method. It would be interesting to propose a rank method for extraction of keywords using wikilinks.

References

[1] Davison, B.D.: Topical locality in the web. In: Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR 2000). pp. 272–279 (2000)

[2] Garcia, R.D., Rensing, C., Steinmetz, R.: Automatic acquisition of taxonomies in different languages from multiple wikipedia versions. In: Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW '11) (2011)

[3] Greenwood, M.A., Stevenson, M.: Improving semi-supervised acquisition of relation extraction patterns. In: Proceedings of the Workshop on Information Extraction Beyond The Document (IEBeyondDoc-06). pp. 29–35 (2006)

[4] Hu, L., Wang, X., Zhang, M., Li, J., Li, X., Shao, C., Tang, J., Liu, Y.: Learning topic hierarchies for wikipedia categories. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (2015)

[5] Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL-03). vol. 1, pp. 423–430 (2003)

[6] Leacock, C., Chodorow, M.: WordNet: An Electronic Lexical Database, chap. Chapter 11: Combining Local Context and WordNet Similarity for Word Sense Identification, p. 265283. MIT Press, Cambridge, MA (1998)

[7] Lin, D., Pantel, P.: Discovery of inference rules for question-answering. Journal Natural Language Engineering pp. 343–360 (2001)

[8] Lin, W.P., Snover, M., Ji, H.: Unsupervised language-independent name translation mining from wikipedia infoboxes. In: Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing. pp. 43–52 (2011)

[9] Makris, C., Plegas, Y., Theodoridis, E.: Improved text annotation with wikipedia entities. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing. pp. 288–295 (2013)

[10] Mihalcea, R.: Using wikipedia for automatic word sense disambiguation. In: Proceedings of NAACL HLT 2007. pp. 196–203 (2007)

[11] Nakayama, K., Hara, T., Nishio, S.: A thesaurus construction method from large scale web dictionaries. In: 21st International Conference on Advanced Information Networking and Applications. pp. 932–939 (2007)

[12] Nivre, J., de Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., Zeman, D.: Universal dependencies v1: A multilingual treebank collection. In: LREC 2016 (2016)

[13] Stevenson, M., Greenwood, M.A.: Dependency pattern models for information extraction. Dependency Pattern Models for Information Extraction 7(13) (2009)

[14] Sudo, K., Sekine, S., Grishman, R.: Automatic pattern acquisition for japanese information extraction. In: Proceedings of the Human Language Technology Conference (HLT-2001) (2001)

[15] Szpektor, I., Tanev, H., Dagan, I., Coppola, B.: Scaling web-based acquisition of entailment relation. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. pp. 41–48 (2004)
[16] Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. p. 133138 (1994)
[17] Yangarber, R.: Counter-training in discovery of semantic patterns. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. No. 343–350 (2003)
[18] Zesch, T., Gurevych, I.: Analysis of the wikipedia category graph for nlp applications. In: Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007). pp. 1–8 (2007)