Close

Web of Science™
Page 1 (Records 1 -- 50)
[1]

Print

Record 1 of 50

Title: Semantic Sentiment Analysis of Arabic Texts

Author(s): Alowaidi, S (Alowaidi, Sana); Saleh, M (Saleh, Mustafa); Abulnaja, O (Abulnaja, Osama)

Source: INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS **Volume:** 8 **Issue:** 2 **Pages:** 256-262 **Published:** FEB 2017

Abstract: Twitter considered as a rich resource to collect people's opinions in different domains and attracted researchers to develop an automatic Sentiment Analysis (SA) model for tweets. In this work, a semantic Arabic Twitter Sentiment Analysis (ATSA) model is developed based on supervised machine learning (ML) approaches and semantic analysis. Most of the existing Arabic SA approaches represent tweets based on the bag-of-words (BoW) model. The main limitation of this model is that it is semantically weak; where words considered as independent features and ignore the semantic associations between them. As a result, synonymous words that appear in two tweets are represented as different independent features. To overcome this limitation, this work proposes enriching the tweets representation with concepts utilizing Arabic WordNet (AWN) as an external knowledge base. In addition, different concepts representation approaches are developed and evaluated with naive Bayes (NB) and support vector machine (SVM) ML classifiers on an Arabic Twitter dataset. The experimental results indicate that using concepts features improves the performance of the ATSA model compared with the basic BoW representation. The improvement reached 4.48% with the SVM classifier and 5.78% with the NB classifier.

Accession Number: WOS:000397411100034

ISSN: 2158-107X

eISSN: 2156-5570

Record 2 of 50

Title: Accurate Identification of Ontology Alignments at Different Granularity Levels

Author(s): Hu, XC (Hu, Xiaocao); Feng, ZY (Feng, Zhiyong); Chen, SZ (Chen, Shizhan); Huang, KM (Huang, Keman); Li, JQ (Li, Jianqiang); Zhou, MC (Zhou, Mengchu)

Source: IEEE ACCESS **Volume:** 5 **Pages:** 105-120 **DOI:** 10.1109/ACCESS.016.2614759 **Published:** 2017

Abstract: As more and more ontologies are defined with different terms, ontology matching plays a crucial role in addressing the semantic heterogeneity problem in many different disciplines. Many efforts have been made to discover correspondences among terms in different ontologies. Most studies directly match two ontologies by utilizing terminological and structural methods that rely on ontologies themselves only. However, the decentralized characteristic of ontologies raises the uncertainty in ontology matching. To address this problem, we propose a four-stage ontology matching framework (FOMF) to enhance ontology matching performance. It is built upon the commonly accepted claim that an external comprehensive knowledge base can be used as a semantic bridge between domain ontologies for ontology matching. First, FOMF semantically maps domain ontologies to a knowledge base and then produces different types of alignments, including equivalence, subclass, sameas, and instance alignments. Similarities between two domain ontologies are next employed to enhance the equivalence and sameas alignments discovery. Finally, based on acquired alignments, inferred alignments are deduced to guarantee the completeness of matching results. Our experimental results show the superiority of the proposed method over the existing ones.

Accession Number: WOS:000395870000009

ISSN: 2169-3536

Record 3 of 50

Title: Decoding the Semantic Content of Natural Movies from Human Brain Activity

Author(s): Huth, AG (Huth, Alexander G.); Lee, T (Lee, Tyler); Nishimoto, S (Nishimoto, Shinji); Bilenko, NY (Bilenko, Natalia Y.); Vu, AT (Vu, An T.); Gallant, JL (Gallant, Jack L.)

Source: FRONTIERS IN SYSTEMS NEUROSCIENCE **Volume:** 10 **Article Number:** 81 **DOI:** 10.3389/fnsys.2016.00081 **Published:** OCT 7 2016

Abstract: One crucial test for any quantitative model of the brain is to show that the model can be used to accurately decode information from evoked brain activity. Several recent neuroimaging studies have decoded the structure or semantic content of static visual images from human brain activity. Here we present a decoding algorithm that makes it possible to decode detailed information about the object and action categories present in natural movies from human brain activity signals measured by functional MRI. Decoding is accomplished using a hierarchical logistic regression (HLR) model that is based on labels that were manually assigned from the WordNet semantic taxonomy. This model makes it possible to simultaneously decode information about both specific and general categories, while respecting the relationships between them. Our results show that we can decode the presence of many object and action categories from averaged blood-oxygen level-dependent (BOLD) responses with a high degree of accuracy (area under the ROC curve > 0.9). Furthermore, we used this framework to test whether semantic relationships defined in the WordNet taxonomy are represented the same way in the human brain. This analysis showed that hierarchical relationships between general categories and atypical examples, such as organism and plant, did not seem to be reflected in representations measured by BOLD fMRI.

Accession Number: WOS:000384803200001

PubMed ID: 27781035

ISSN: 1662-5137

Record 4 of 50

Title: Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus

Author(s): Atoum, I (Atoum, Issa); Otoom, A (Otoom, Ahmed)

Source: INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS **Volume:** 7 **Issue:** 9 **Pages:** 124-130 **Published:** SEP 2016

Abstract: Text similarity plays an important role in natural language processing tasks such as answering questions and summarizing text. At present, state-of-the-art text similarity algorithms rely on inefficient word pairings and/or knowledge derived from large corpora such as Wikipedia. This article evaluates previous word similarity measures on benchmark datasets and then uses a hybrid word similarity in a novel text similarity measure (TSM). The proposed TSM is based on information content and WordNet semantic relations. TSM includes exact word match, the length of both sentences in a pair, and the maximum similarity between one word and the compared text. Compared with other well-known measures, results of TSM are surpassing or comparable with the best algorithms in the literature.

Accession Number: WOS:000391110500018

Author Identifiers:

Author	ResearcherID Number	ORCID Number
Atoum, Dr. Issa	O-9388-2014	0000-0002-2160-3615

ISSN: 2158-107X

eISSN: 2156-5570

Record 5 of 50

Title: A Novel Information Retrieval Approach using Query Expansion and Spectral-based

Author(s): Alnofaie, S (Alnofaie, Sara); Dahab, M (Dahab, Mohammed); Kamal, M (Kamal, Mahmoud)

Source: INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS **Volume:** 7 **Issue:** 9 **Pages:** 364-373 **Published:** SEP 2016

Abstract: Most of the information retrieval (IR) models rank the documents by computing a score using only the lexicographical query terms or frequency information of the query terms in the document. These models have a limitation as they does not consider the terms proximity in the document or the term-mismatch or both of the two. The terms proximity information is an important factor that determines the relatedness of the document to the query. The ranking functions of the Spectral-Based Information Retrieval Model (SBIRM) consider the query terms frequency and proximity in the document by comparing the signals of the query terms in the spectral domain instead of the spatial domain using Discrete Wavelet Transform (DWT). The query expansion (QE) approaches are used to overcome the word-mismatch problem by adding terms to query, which have related meaning with the query. The QE approaches are divided to statistical approach Kullback-Leibler divergence (KLD) and semantic approach P-WNET that uses WordNet. These approaches enhance the performance. Based on the foregoing considerations, the objective of this research is to build an efficient QESBIRM that combines QE and proximity SBIRM by implementing the SBIRM using the DWT and KLD or P-WNET. The experiments conducted to test and evaluate the QESBIRM using Text Retrieval Conference (TREC) dataset. The result shows that the SBIRM with the KLD or P-WNET model outperform the SBIRM model in precision (P@), R-precision, Geometric Mean Average Precision (GMAP) and Mean Average Precision (MAP).

Accession Number: WOS:000391110500051

ISSN: 2158-107X

eISSN: 2156-5570

Record 6 of 50

Title: MMO: Multiply-Minus-One Rule for Detecting & Ranking Positive and Negative Opinion

Author(s): Saqib, SM (Saqib, Sheikh Muhammad); Kundi, FM (Kundi, Fazal Masud)

Source: INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS **Volume:** 7 **Issue:** 5 **Pages:** 122-127 **Published:** MAY 2016

Abstract: Hit and hot issue about reviews of any product is sentiment classification. Not only manufacturing company of the reviewed product takes decision about its quality, but the customers' purchase of the product is also based on the reviews. Instead of reading all the reviews one by one, different works have been done to classify them as negative or positive with preprocessing. Suppose from 1000 reviews, there are 300 negative and 700 are positive. As a whole it is positive. Company and customer may not be satisfied with this sentiment orientation. For companies, negative reviews should be separated with respect to different aspects and features, so companies can enhance the features of the product. There is also a lot of work on aspect extraction, and then aspect based sentiment analysis. While on the other hand, users want the most positive reviews and the most negative reviews, then they can decide purchasing a certain product. To consider the issue from users' perspective, authors suggest a method Multiply-Minus-One (MMO) which can evaluate each review and find scores based on positive, negative, intensifiers and negation words using WordNet Dictionary. Experiments on 4 types of datasets of product reviews show that this method can achieve 86%, 83%, 83% and 85% precision performance.

Accession Number: WOS:000377221700019

ISSN: 2158-107X

eISSN: 2156-5570

Record 7 of 50

Title: Word Sense Disambiguation Approach for Arabic Text

Author(s): Bouhriz, N (Bouhriz, Nadia); Benabbou, F (Benabbou, Faouzia); Ben Lahmar, E (Ben Lahmar, El Habib)

Source: INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS **Volume:** 7 **Issue:** 4 **Pages:** 381-385 **Published:** APR 2016

Abstract: Word Sense Disambiguation (WSD) consists of identifying the correct sense of an ambiguous word occurring in a given context. Most of Arabic WSD systems are based generally on the information extracted from the local context of the word to be disambiguated. This information is not usually sufficient for a best disambiguation. To overcome this limit, we propose an approach that takes into consideration, in addition to the local context, the global context too extracted from the full text. More particularly, the sense attributed to an ambiguous word is the one of which semantic proximity is more close both to its local and global context. The experiments show that the proposed system achieved an accuracy of 74%.

Accession Number: WOS:000377221300051

ISSN: 2158-107X

eISSN: 2156-5570

Record 8 of 50

Title: Integrating Semantic Features for Enhancing Arabic Named Entity Recognition

Author(s): Alsayadi, HA (Alsayadi, Hamzah A.); ElKorany, AM (ElKorany, Abeer M.)

Source: INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS **Volume:** 7 **Issue:** 3 **Pages:** 128-136 **Published:** MAR 2016

Abstract: Named Entity Recognition (NER) is currently an essential research area that supports many tasks in NLP. Its goal is to find a solution to boost accurately the named entities identification. This paper presents an integrated semantic-based Machine learning (ML) model for Arabic Named Entity Recognition (ANER) problem. The basic idea of that model is to combine several linguistic features and to utilize syntactic dependencies to infer semantic relations between named entities. The proposed model focused on recognizing three types of named entities: person, organization and location. Accordingly, it combines internal features that represented linguistic features as well as external features that represent the semantic of relations between the three named entities to enhance the accuracy of recognizing them using external knowledge source such as Arabic WordNet ontology (ANW). We introduced both features to CRF classifier, which are effective for ANER. Experimental results show that this approach can achieve an overall F-measure around 87.86% and 84.72% for ANERCorp and ALTEC datasets respectively.

Accession Number: WOS:000377220600018

ISSN: 2158-107X

eISSN: 2156-5570

Record 9 of 50

Title: Text Mining the History of Medicine

Author(s): Thompson, P (Thompson, Paul); Batista-Navarro, RT (Batista-Navarro, Riza Theresa); Kontonatsios, G (Kontonatsios, Georgios); Carter, J (Carter, Jacob); Toon, E (Toon, Elizabeth); McNaught, J (McNaught, John); Timmermann, C (Timmermann, Carsten); Worboys, M (Worboys, Michael); Ananiadou, S (Ananiadou, Sophia)

Source: PLOS ONE **Volume:** 11 **Issue:** 1 **Article Number:** e0144717 **DOI:** 10.1371/journal.pone.0144717 **Published:** JAN 6 2016

Abstract: Historical text archives constitute a rich and diverse source of information, which is becoming increasingly readily accessible, due to large-scale digitisation efforts. However, it can be difficult for researchers to explore and search such large volumes of data in an efficient manner. Text mining (TM) methods can help, through their ability to recognise various types of semantic information automatically, e.g., instances of concepts (places, medical conditions, drugs, etc.), synonyms/variant forms of concepts, and relationships holding between concepts (which drugs are used to treat which medical conditions, etc.). TM analysis allows search systems to incorporate functionality such as automatic suggestions of synonyms of user-entered query terms, exploration of different concepts mentioned within search results or isolation of documents in which concepts are related in specific ways. However, applying TM methods to historical text can be challenging, according to differences and evolutions in vocabulary, terminology, language structure and style, compared to more modern text. In this article, we present our efforts to overcome the various challenges faced in the semantic analysis of published historical medical text dating back to the mid 19th century. Firstly, we used evidence from diverse historical medical documents from different periods to develop new resources that provide accounts of the multiple, evolving ways in which concepts, their variants and relationships amongst them may be expressed. These resources were employed to support the development of a modular processing pipeline of TM tools for the robust detection of semantic information in historical medical documents with varying characteristics. We applied the pipeline to two large-scale medical document archives covering wide temporal ranges as the basis for the development of a publicly accessible semantically-oriented search system. The novel resources are available for research purposes, while the processing pipeline and its modules may be used and configured within the Argo TM platform.

Accession Number: WOS:000367805100011

PubMed ID: 26734936

ISSN: 1932-6203

Record 10 of 50

Title: Introducing Idioms in the Galician WordNet: Methods, Problems and Results

Author(s): de la Granja, MA (Alvarez de la Granja, Maria); Clemente, XMG (Gomez Clemente, Xose Maria); Guinovart, XG (Gomez Guinovart, Xavier)

Source: OPEN LINGUISTICS **Volume:** 2 **Issue:** 1 **Pages:** 253-286 **DOI:** 10.1515/opli-2016-0012 **Published:** JAN 2016

Abstract: This study describes the introduction of verbal idioms in the Galician language version (Galnet) of the semantic network WordNet; a network that does not traditionally include many phraseological units. To enhance Galnet, a list of 803 Galician verbal idioms was developed to then review each of them individually and assess whether they could be introduced in an existing WordNet synset (a group of synonyms expressing the same concept) or not. Of those 803 idioms, 490 (61%) could be included in this network. Besides, Galnet was enlarged with 750 extra verbal idioms, most of them synonyms or variants of the former. In this study, we present the working methodology for the experiment and an analysis of the results, to help understand the most important problems found when trying to introduce idioms in Galnet. We also discuss the reasons preventing the inclusion of some expressions, and the criteria used to introduce the idioms that finally made it into the network.

Accession Number: WOS:000397102300005

ISSN: 2300-9969

Record 11 of 50

Title: Polysemy and homonymy of idioms in a lexical model

Author(s): Pause, MS (Pause, Marie-Sophie); Sikora, D (Sikora, Dorota)

Edited by: Neveu F; Bergounioux G; Cote MH; Fournier JM; Hriba L; Prevost S

Source: 5E CONGRES MONDIAL DE LINGUISTIQUE FRANCAISE **Book Series:** SHS Web of Conferences **Volume:** 27 **Article Number:** 05012 **DOI:** 10.1051/shsconf/20162705012 **Published:** 2016

Abstract: The present paper focuses on a model of linguistic analysis that enables lexicographers to draw a clear distinction between polysemy and homonymy of idioms. Indeed, researchers' growing interest in phraseological units leads to give more room to idioms in lexicographic description (see for instance WordNet[1], FrameNet[2]). We therefore believe that, few dictionaries-if any-systematically account for their polysemy. The lexical database Reseau Lexical du Francais (French Lexical Network, henceforth RL-fr) currently developed in the laboratory ATILF -CNRS (UMR 7118)[3] presently encodes about 3000 French idioms, with a systematic treatment of their polysemous structure. Building on this large-scale lexicographic work based on theoretical et methodological principles of Explanatory and Combinatorial Lexicography (Mel'cuk et al. 1995), we examine the case of the verbal idiom marcher sur la tete (lit. 'to walk on the head'). It conveys three senses to be described in a dictionary: 'to dominate', 'to inflict a crushing defeat (in a competition)', and 'to behave irrationally'. Thus, the question that arises is whether there is one polysemous idiom marcher sur la tete or if we should rather consider it a case of homonymy. Semantic analysis can hardly establish any meaning relation between 'to dominate' and 'to inflict a crushing defeat' on the one side, and 'to behave irrationally' on the other. However, lexicographic treatment of idioms in RL-fr takes into account the signified as well as the signifier: the form of the later is modelized as its lexico-syntactic structure. As a matter of fact, marcher sur la tete reveals two different lexico-syntactic structures, corresponding respectively to the polysemous idiom marcher sur la tete1 with two meanings ('to dominate', 'to inflict a crushing defeat') and its homonym marcher sur la tete2- 'to behave irrationally'. These results lead to conclude that formal properties of idioms, modelized as their lexico-syntactic structure, should systematically be accounted for in lexicographic description. Moreover, given the available data, we suggest that lexico-syntactic structures of idioms correspond to their signifiers.

Accession Number: WOS:000392694300042

Conference Title: 5th World Congress of French Linguistics

Conference Date: JUL 04-08, 2016

Conference Location: Tours, FRANCE

ISSN: 2261-2424

Record 12 of 50

Title: SHAPELEARNER: TOWARDS SHAPE-BASED VISUAL KNOWLEDGE HARVESTING

Author(s): Wang, Z (Wang, Zheng); Liang, T (Liang, Ti)

Edited by: Halounova L; Schindler K; Limpouch A; Pajdla T; Safar V; Mayer H; Elberink SO; Mallet C; Rottensteiner F; Bredif M; Skaloud J; Stilla U

Source: XXIII ISPRS CONGRESS, COMMISSION III **Book Series:** International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences **Volume:** 41 **Issue:** B3 **Pages:** 789-796 **DOI:** 10.5194/isprsarchives-XLI-B3-789-2016 **Published:** 2016

Abstract: The explosion of images on the Web has led to a number of efforts to organize images semantically and compile collections of visual knowledge. While there has been enormous progress on categorizing entire images or bounding boxes, only few studies have targeted fine-grained image understanding at the level of specific shape contours. For example, given an image of a cat, we would like a system to not merely recognize the existence of a cat, but also to distinguish between the cat's legs, head, tail, and so on. In this paper, we present ShapeLearner, a system that acquires such visual knowledge about object shapes and their parts. ShapeLearner jointly learns this knowledge from sets of segmented images. The space of label and segmentation hypotheses is pruned and then evaluated using Integer Linear Programming. ShapeLearner places the resulting knowledge in a semantic taxonomy based on WordNet and is able to exploit this hierarchy in order to analyze new kinds of objects that it has not observed before. We conduct experiments using a variety of shape classes from several representative categories and demonstrate the accuracy and robustness of our method.

Accession Number: WOS:000392743800115

Conference Title: 23rd Congress of the International-Society-for-Photogrammetry-and-Remote-Sensing (ISPRS)

Conference Date: JUL 12-19, 2016

Conference Location: Prague, CZECH REPUBLIC

Conference Sponsors: Int Soc Photogrammetry & Remote Sensing

ISSN: 2194-9034

Record 13 of 50

Title: Game with a Purpose for Mappings Verification

Author(s): Boinski, T (Boinski, Tomasz)

Edited by: Ganzha M; Maciaszek L; Paprzycki M

Source: PROCEEDINGS OF THE 2016 FEDERATED CONFERENCE ON COMPUTER SCIENCE AND INFORMATION SYSTEMS (FEDCSIS) **Book Series:** ACSIS-Annals of Computer Science and Information Systems **Volume:** 8 **Pages:** 405-409 **DOI:** 10.15439/2016F172 **Published:** 2016

Abstract: Mappings verification is a laborious task. The paper presents a Game with a Purpose based system for verification of automatically generated mappings. General description of idea standing behind the games with the purpose is given. Description of TGame system, a 2D platform mobile game with verification process included in the gameplay, is provided. Additional mechanisms for anti-cheating, increasing player's motivation and gathering feedback are also presented. Example of the system usage for verification of mappings between WordNet synsets and Wikipedia articles is presented. The evaluation of proposed solution and future work is

also described.
Accession Number: WOS:000392436600059
Conference Title: Federated Conference on Computer Science and Information Systems (FedCSIS)
Conference Date: SEP 11-14, 2016
Conference Location: Gdansk, POLAND
Conference Sponsors: PTI, IEEE
ISSN: 2300-5963
ISBN: 978-8-3608-1090-3

Record 14 of 50
Title: Recommending design patterns using task-based conceptual features
Author(s): Laosen, N (Laosen, Nasith); Sanyawong, N (Sanyawong, Nuttapon); Nantajeewarawat, E (Nantajeewarawat, Ekawit)
Source: MAEJO INTERNATIONAL JOURNAL OF SCIENCE AND TECHNOLOGY **Volume:** 10 **Issue:** 1 **Pages:** 113-126 **DOI:** 10.14456/mijst.2016.11 **Published:** JAN-APR 2016
Abstract: Task-based conceptual features (TCFs) represent human knowledge concerning intentions and/or characteristics of design tasks to which a design pattern is applicable. They provide a bridge connecting the usage of a design pattern and the characteristics of a design problem. A method for recommending appropriate design patterns based on TCFs is presented. From grammatical relations between words generated from the textual description of an input design problem, problem keywords are extracted. The obtained problem keywords are matched with clue words of each TCF in order to construct a feature vector representing the input problem. Based on the similarity between the feature vector representing the problem and the TCF-based vector representing each design pattern, design patterns are ranked and recommended. The method is evaluated on a collection of 24 input design problems. The evaluation results show that when the first-level hypernyms and hyponyms obtained from the WordNet ontology are employed for word matching and an appropriate penalty score is assigned to them, design patterns recommended in the top three ranks include the correct design patterns for all 24 problems.
Accession Number: WOS:000383727500008
ISSN: 1905-7873

Record 15 of 50
Title: Automatic Wordnet Development for Low-Resource Languages using Cross-Lingual WSD
Author(s): Taghizadeh, N (Taghizadeh, Nasrin); Faili, H (Faili, Hesham)
Source: JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH **Volume:** 56 **Pages:** 61-87 **Published:** 2016
Abstract: Wordnet is an effective resource in natural language processing and information retrieval , especially for semantic processing and meaning related tasks. So far wordnet has been constructed in many languages. However, automatic development of wordnet for low-resource languages has not been studied well. In this paper an Expectation-Maximization algorithm is used to train high quality and large scale wordnet for resource-poor languages. The proposed method benefits from cross-lingual word sense disambiguation and develops a wordnet just using a bilingual dictionary and a monolingual corpus. The proposed method has been executed on Persian as a resource-poor language and the resulting wordnet has been evaluated through several experiments. Results show that the induced wordnet has a precision of 90% and recall of 35%.
Accession Number: WOS:000380243600001
ISSN: 1076-9757
eISSN: 1943-5037

Record 16 of 50
Title: Effectiveness of Automatic Translations for Cross-Lingual Ontology Mapping
Author(s): Abu Helou, M (Abu Helou, Mamoun); Palmonari, M (Palmonari, Matteo); Jarrar, M (Jarrar, Mustafa)
Source: JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH **Volume:** 55 **Pages:** 165-208 **Published:** 2016
Abstract: Accessing or integrating data lexicalized in different languages is a challenge. Multilingual lexical resources play a fundamental role in reducing the language barriers to map concepts lexicalized in different languages. In this paper we present a large-scale study on the effectiveness of automatic translations to support two key cross-lingual ontology mapping tasks: the retrieval of candidate matches and the selection of the correct matches for inclusion in the final alignment. We conduct our experiments using four different large gold standards, each one consisting of a pair of mapped wordnets, to cover four different families of languages. We categorize concepts based on their lexicalization (type of words, synonym richness, position in a subconcept graph) and analyze their distributions in the gold standards. Leveraging this categorization, we measure several aspects of translation effectiveness, such as word-translation correctness, word sense coverage, synset and synonym coverage. Finally, we thoroughly discuss several findings of our study, which we believe are helpful for the design of more sophisticated cross-lingual mapping algorithms.
Accession Number: WOS:000370578500001
Author Identifiers:

Author	ResearcherID Number	ORCID Number
PALMONARI, MATTEO		0000-0002-1801-5118

ISSN: 1076-9757
eISSN: 1943-5037

Record 17 of 50
Title: Discovering Fuzzy Synsets from the Redundancy across several Dictionaries
Author(s): Santos, F (Santos, Fabio); Oliveira, HG (Oliveira, Hugo Goncalo)
Source: LINGUAMATICA **Volume:** 7 **Issue:** 2 **Pages:** 3-17 **Published:** DEC 2015
Abstract: In a wordnet, concepts are typically represented as groups of words, commonly known as synsets, and each membership of a word to a synset denotes a different sense of that word. However, since word senses are complex entities, without well-defined boundaries, we suggest to handle them less artificially, by representing them as fuzzy objects, where each word has its membership degree, which can be related to the confidence on using the word to denote the concept conveyed by the synset. We thus propose an approach to discover synsets from a synonymy network, ideally redundant and extracted from several broad-coverage sources. The more synonymy relations there are between two words, the higher the confidence on the semantic equivalence of at least one of their senses. The proposed approach was applied to a network extracted from three Portuguese dictionaries and resulted in a large set of fuzzy synsets. Besides describing this approach and illustrating its results, we rely on three evaluations - comparison against a handcrafted Portuguese thesaurus; comparison against the results of a previous approach with a similar goal; and manual evaluation - to believe that our outcomes are positive and that, in the future, they might my expanded by exploring additional synonymy sources.
Accession Number: WOS:000371641800002
ISSN: 1647-0818

Record 18 of 50
Title: A Gloss Composition and Context Clustering Based Distributed Word Sense Representation Model
Author(s): Chen, T (Chen, Tao); Xu, RF (Xu, Ruifeng); He, YL (He, Yulan); Wang, X (Wang, Xuan)

Source: ENTROPY **Volume:** 17 **Issue:** 9 **Pages:** 6007-6024 **DOI:** 10.3390/e17096007 **Published:** SEP 2015

Abstract: In recent years, there has been an increasing interest in learning a distributed representation of word sense. Traditional context clustering based models usually require careful tuning of model parameters, and typically perform worse on infrequent word senses. This paper presents a novel approach which addresses these limitations by first initializing the word sense embeddings through learning sentence-level embeddings from WordNet glosses using a convolutional neural networks. The initialized word sense embeddings are used by a context clustering based model to generate the distributed representations of word senses. Our learned representations outperform the publicly available embeddings on half of the metrics in the word similarity task, 6 out of 13 sub tasks in the analogical reasoning task, and gives the best overall accuracy in the word sense effect classification task, which shows the effectiveness of our proposed distributed distribution learning model.

Accession Number: WOS:000362513400005

Author Identifiers:

Author	ResearcherID Number	ORCID Number
He, Yulan		0000-0003-3948-5845

ISSN: 1099-4300

Record 19 of 50

Title: Tuning a Semantic Relatedness Algorithm using a Multiscale Approach

Author(s): Leal, JP (Leal, Jose Paulo); Costa, T (Costa, Teresa)

Source: COMPUTER SCIENCE AND INFORMATION SYSTEMS **Volume:** 12 **Issue:** 2 **Pages:** 635-654 **DOI:** 10.2298/CSIS140905020L **Published:** JUN 2015

Abstract: The research presented in this paper builds on previous work that lead to the definition of a family of semantic relatedness algorithms. These algorithms depend on a semantic graph and on a set of weights assigned to each type of arcs in the graph. The current objective of this research is to automatically tune the weights for a given graph in order to increase the proximity quality. The quality of a semantic relatedness method is usually measured against a benchmark data set. The results produced by a method are compared with those on the benchmark using a nonparametric measure of statistical dependence, such as the Spearman's rank correlation coefficient. The presented methodology works the other way round and uses this correlation coefficient to tune the proximity weights. The tuning process is controlled by a genetic algorithm using the Spearman's rank correlation coefficient as fitness function. This algorithm has its own set of parameters which also need to be tuned. Bootstrapping is a statistical method for generating samples that is used in this methodology to enable a large number of repetitions of a genetic algorithm, exploring the results of alternative parameter settings. This approach raises several technical challenges due to its computational complexity. This paper provides details on techniques used to speedup the process. The proposed approach was validated with the Word Net 2.1 and the Word Sim-353 data set. Several ranges of parameter values were tested and the obtained results are better than the state of the art methods for computing semantic relatedness using the Word Net 2.1, with the advantage of not requiring any domain knowledge of the semantic graph.

Accession Number: WOS:000359611300016

Author Identifiers:

Author	ResearcherID Number	ORCID Number
Leal, Jose Paulo		0000-0002-8409-0300

ISSN: 1820-0214

Record 20 of 50

Title: Two Stage Optimization Model to Semantic Service Discovery

Author(s): Ganapathy, G (Ganapathy, Gopinath); Surianarayanan, C (Surianarayanan, Chellammal)

Edited by: Ao SI; Douglas C; Grundfest WS; Burgstone J

Source: WORLD CONGRESS ON ENGINEERING AND COMPUTER SCIENCE, WCECS 2015, VOL I **Book Series:** Lecture Notes in Engineering and Computer Science **Pages:** 439-443 **Published:** 2015

Abstract: Discovering appropriate services quickly for dynamic service composition is a challenging issue. Clustering technique partitions the available services into clusters of similar services. During discovery of matched services for a query, semantic matching of service capabilities is performed only to a particular cluster which is most relevant to the query and other clusters are ignored as irrelevant. Thus clustering improves the performance of semantic discovery by eliminating irrelevancy. In one of our previous research work, two similarity models, one for computing similarity between services(called Output Similarity Model) while clustering them and the other(called Total Similarity Model) for finding matched services for a given query using clusters along with selection of similarity threshold and recommendation of complete linkage criterion for computing inter-cluster distance are proposed for service discovery using hierarchical agglomerative clustering. As an extension of our previous work, in this paper, an experimental evaluation has been performed to analyze the performance of OSM in regard to effective removal of irrelevancy and the strength of prioritizing parameters during discovery. Further, the clustering solutions obtained using Output Similarity Model are compared with those produced by standard methods such as syntactic similarity and WordNet similarity based, methods. Though clustering improves the performance of discovery by eliminating irrelevant clusters, still is required to employ semantic matching to the services present in the relevant cluster. This involves invoking semantic reasoning during querying. To resolve this limitation, after clustering, an indexing technique is suggested to the resulting clustering solution. With this model, the invoking of semantic reasoning is completely eliminated.

Accession Number: WOS:000380591600084

Conference Title: World Congress on Engineering and Computer Science

Conference Date: OCT 21-23, 2015

Conference Location: San Francisco, CA

Conference Sponsors: Int Assoc Engn

ISSN: 2078-0958

ISBN: 978-988-19253-6-7

Record 21 of 50

Title: Feature and Sentiment based Linked Instance RDF Data towards Ontology based Review Categorization

Author(s): Santosh, DT (Santosh, D. Teja); Vardhan, BV (Vardhan, B. Vishnu)

Edited by: Ao SI; Gelman L; Hukins DWL; Hunter A; Korsunsky AM

Source: WORLD CONGRESS ON ENGINEERING, WCE 2015, VOL I **Book Series:** Lecture Notes in Engineering and Computer Science **Pages:** 493-497 **Published:** 2015

Abstract: Online reviews have a potential impact on the green customer who wants to purchase or consume the product through e-commerce. Online reviews contain features which are useful for the analysis in opinion mining. Most of the today's systems work on the summarization of the features taking the average features and their sentiments leading to structured review information. Often the context of surrounding feature is undermined which helps while classifying the sentiment of the review. In web 3.0 machine interpretable Resource Description Framework (RDF) were introduced which helps in structuring these unstructured reviews in the form of features and sentiments obtained from traditional preprocessing and extraction techniques. The context data also supports for future ontology based analysis taking support of Wordnet lexical database for word sense disambiguation and Sentiwordnet scores used for sentiment word extraction. Many popular RDF vocabularies are helpful in the creation of such machine processable data. In the future work, such instance RDF data will be used in the OWL Ontology to reason the data to clearly identify the features and sentiments against the applied data set. These results are sent back to the interface as corresponding {feature, sentiment} pair so that reviews are filtered clearly and helps in satisfying the feature set of the customer.

Accession Number: WOS:000380592400097
Conference Title: World Congress on Engineering (WCE 2015)
Conference Date: JUL 01-03, 2015
Conference Location: Imperial Coll, London, ENGLAND
Conference Sponsors: Int Assoc Engn
Conference Host: Imperial Coll
ISSN: 2078-0958
ISBN: 978-988-19253-4-3

Record 22 of 50
Title: Verb Sense Annotation in News Texts in the CSTNews Corpus
Author(s): Cabezudo, MAS (Sobrevilla Cabezudo, Marco Antonio); Maziero, EG (Maziero, Erick Galani); Souza, JWD (da Cruz Souza, Jackson Wilke); Dias, MD (Dias, Myrcio de Souza); Cardoso, PCF (Figueira Cardoso, Paula Christina); Balage, PP (Balage Filho, Pedro Paulo); Agostini, V (Agostini, Veronica); Nobrega, FAA (Asevedo Nobrega, Fernando Antonio); de Barros, CD (de Barros, Claudia Dias); Di Felippo, A (Di Felippo, Ariani); Pardo, TAS (Salgueiro Pardo, Thiago Alexandre)
Source: REVISTA DE ESTUDOS DA LINGUAGEM **Volume:** 23 **Issue:** 3 **Pages:** 797-832 **Published:** 2015
Abstract: One of the hardest problems in Natural Language Processing (NLP) is the lexical ambiguity, as words may express different senses depending on the context in which they occur. In NLP, Word Sense Disambiguation (WSD) is the task that aims at determining the proper meaning of a word in its context. In this task, the use of a sense annotated corpus is useful because this computational linguistic resource enables further study of the ambiguity phenomenon and the development and evaluation of WSD methods. This paper describes the verb sense annotation process in news texts in the CSTNews corpus, using Princeton WordNet as sense repository. Besides detailing the annotation process and its results, the contributions of this work include the availability of a linguistic resource that may be the basis for future research in WSD for Portuguese.
Accession Number: WOS:000371826300008
Author Identifiers:

Author	ResearcherID Number	ORCID Number
Inst Cien Matematicas Computacao, ICMC/USP	D-8320-2017	

ISSN: 0104-0588
eISSN: 2237-2083

Record 23 of 50
Title: OBTAINING FEATURE- AND SENTIMENT-BASED LINKED INSTANCE RDF DATA FROM UNSTRUCTURED REVIEWS USING ONTOLOGY-BASED MACHINE LEARNING
Author(s): Santosh, DT (Santosh, D. Teja); Vardhan, BV (Vardhan, B. Vishnu)
Source: INTERNATIONAL JOURNAL OF TECHNOLOGY **Volume:** 6 **Issue:** 2 **Pages:** 198-206 **Published:** 2015
Abstract: Online reviews have a profound impact on the customer or "newbie" who wishes to purchase or consume a product via Web 2.0 e-commerce. Online reviews contain features that form half of the analysis in opinion mining. Most of today's systems work on the basis of summarization, looking at the average obtained features and their sentiments, leading to structured review information being generated. Often, the context surrounding a feature, which helps the sentiment of the review to be classified clearly, is overlooked. The Web 3.0-based machine interpretable Resource Description Framework (RDF) can be used to structure these unstructured reviews into features and sentiments, which are obtained via traditional preprocessing and extraction techniques. Here, data about the context is also provided for future ontology-based analysis, with support from the WordNet lexical database for word sense disambiguation and SentiWordNet scores for sentiment word extraction. Many popular RDF vocabularies are helpful for obtaining such machine-processable data. This work forms the basis for creating/upgrading the (available) OWL Ontology that can be used as a structured data model with rich semantics for supervised machine learning. With this method, the classified sentiment categories are validated in relation to precise sentiments and are sent back to the interface in corresponding "feature/sentiment" pairs so that reviews are filtered clearly, which helps to satisfy the feature set of the customer.
Accession Number: WOS:000363011300011
ISSN: 2086-9614
eISSN: 2087-2100

Record 24 of 50
Title: Word vs. Class-BasedWord Sense Disambiguation
Author(s): Izquierdo, R (Izquierdo, Ruben); Suarez, A (Suarez, Armando); Rigau, G (Rigau, German)
Source: JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH **Volume:** 54 **Pages:** 83-122 **Published:** 2015
Abstract: As empirically demonstrated by the Word Sense Disambiguation (WSD) tasks of the last SensEval/SemEval exercises, assigning the appropriate meaning to words in context has resisted all attempts to be successfully addressed. Many authors argue that one possible reason could be the use of inappropriate sets of word meanings. In particular, WordNet has been used as a de-facto standard repository of word meanings in most of these tasks. Thus, instead of using the word senses defined in WordNet, some approaches have derived semantic classes representing groups of word senses. However, the meanings represented by WordNet have been only used for WSD at a very fine-grained sense level or at a very coarse-grained semantic class level (also called SuperSenses). We suspect that an appropriate level of abstraction could be on between both levels. The contributions of this paper are manifold. First, we propose a simple method to automatically derive semantic classes at intermediate levels of abstraction covering all nominal and verbal WordNet meanings. Second, we empirically demonstrate that our automatically derived semantic classes outperform classical approaches based on word senses and more coarse-grained sense groupings. Third, we also demonstrate that our supervised WSD system benefits from using these new semantic classes as additional semantic features while reducing the amount of training examples. Finally, we also demonstrate the robustness of our supervised semantic class-based WSD system when tested on out of domain corpus.
Accession Number: WOS:000365177800002
ISSN: 1076-9757
eISSN: 1943-5037

Record 25 of 50
Title: Fast Distributed Dynamics of Semantic Networks via Social Media
Author(s): Carrillo, F (Carrillo, Facundo); Cecchi, GA (Cecchi, Guillermo A.); Sigman, M (Sigman, Mariano); Slezak, DF (Fernandez Slezak, Diego)
Source: COMPUTATIONAL INTELLIGENCE AND NEUROSCIENCE **Article Number:** 712835 **DOI:** 10.1155/2015/712835 **Published:** 2015
Abstract: We investigate the dynamics of semantic organization using social media, a collective expression of human thought. We propose a novel, time-dependent semantic similarity measure (TSS), based on the social network Twitter. We show that TSS is consistent with static measures of similarity but provides high temporal resolution for the identification of real-world events and induced changes in the distributed structure of semantic relationships across the entire lexicon. Using TSS, we measured the evolution of a concept and its movement along the semantic neighborhood, driven by specific news/events. Finally, we showed that particular events may trigger a temporary reorganization of elements in the semantic network.
Accession Number: WOS:000355484600001
ISSN: 1687-5265

eISSN: 1687-5273

Record 26 of 50

Title: Anatomical Entity Recognition with a Hierarchical Framework Augmented by External Resources

Author(s): Xu, Y (Xu, Yan); Hua, J (Hua, Ji); Ni, ZH (Ni, Zhaoheng); Chen, QL (Chen, Qinlang); Fan, YB (Fan, Yubo); Ananiadou, S (Ananiadou, Sophia); Chang, EIC (Chang, Eric I-Chao); Tsujii, J (Tsujii, Junichi)

Source: PLOS ONE **Volume:** 9 **Issue:** 10 **Article Number:** e108396 **DOI:** 10.1371/journal.pone.0108396 **Published:** OCT 24 2014

Abstract: References to anatomical entities in medical records consist not only of explicit references to anatomical locations, but also other diverse types of expressions, such as specific diseases, clinical tests, clinical treatments, which constitute implicit references to anatomical entities. In order to identify these implicit anatomical entities, we propose a hierarchical framework, in which two layers of named entity recognizers (NERs) work in a cooperative manner. Each of the NERs is implemented using the Conditional Random Fields (CRF) model, which use a range of external resources to generate features. We constructed a dictionary of anatomical entity expressions by exploiting four existing resources, i.e., UMLS, MeSH, RadLex and BodyPart3D, and supplemented information from two external knowledge bases, i.e., Wikipedia and WordNet, to improve inference of anatomical entities from implicit expressions. Experiments conducted on 300 discharge summaries showed a micro-averaged performance of 0.8509 Precision, 0.7796 Recall and 0.8137 F1 for explicit anatomical entity recognition, and 0.8695 Precision, 0.6893 Recall and 0.7690 F1 for implicit anatomical entity recognition. The use of the hierarchical framework, which combines the recognition of named entities of various types (diseases, clinical tests, treatments) with information embedded in external knowledge bases, resulted in a 5.08% increment in F1. The resources constructed for this research will be made publicly available.

Accession Number: WOS:000343943500005

PubMed ID: 25343498

ISSN: 1932-6203

Record 27 of 50

Title: A Bayesian generative model for learning semantic hierarchies

Author(s): Mittelman, R (Mittelman, Roni); Sun, M (Sun, Min); Kuipers, B (Kuipers, Benjamin); Savarese, S (Savarese, Silvio)

Source: FRONTIERS IN PSYCHOLOGY **Volume:** 5 **Article Number:** 417 **DOI:** 10.3389/fpsyg.2014.00417 **Published:** MAY 20 2014

Abstract: Building fine-grained visual recognition systems that are capable of recognizing tens of thousands of categories, has received much attention in recent years. The well known semantic hierarchical structure of categories and concepts, has been shown to provide a key prior which allows for optimal predictions. The hierarchical organization of various domains and concepts has been subject to extensive research, and led to the development of the Word Net domains hierarchy (Fellbaum, 1998), which was also used to organize the images in the Image Net (Deng et al., 2009) dataset, in which the category count approaches the human capacity. Still, for the human visual system, the form of the hierarchy must be discovered with minimal use of supervision or innate knowledge. In this work, we propose a new Bayesian generative model for learning such domain hierarchies, based on semantic input. Our model is motivated by the super-subordinate organization of domain labels and concepts that characterizes WordNet, and accounts for several important challenges: maintaining context information when progressing deeper into the hierarchy, learning a coherent semantic concept for each node, and modeling uncertainty in the perception process.

Accession Number: WOS:000336085600001

PubMed ID: 24904452

ISSN: 1664-1078

Record 28 of 50

Title: Integrating Semantic Information into Multiple Kernels for Protein-Protein Interaction Extraction from Biomedical Literatures

Author(s): Li, LS (Li, Lishuang); Zhang, PP (Zhang, Panpan); Zheng, TF (Zheng, Tianfu); Zhang, HY (Zhang, Hongying); Jiang, ZC (Jiang, Zhenchao); Huang, DG (Huang, Degen)

Source: PLOS ONE **Volume:** 9 **Issue:** 3 **Article Number:** e91898 **DOI:** 10.1371/journal.pone.0091898 **Published:** MAR 12 2014

Abstract: Protein-Protein Interaction (PPI) extraction is an important task in the biomedical information extraction. Presently, many machine learning methods for PPI extraction have achieved promising results. However, the performance is still not satisfactory. One reason is that the semantic resources were basically ignored. In this paper, we propose a multiple-kernel learning-based approach to extract PPIs, combining the feature-based kernel, tree kernel and semantic kernel. Particularly, we extend the shortest path-enclosed tree kernel (SPT) by a dynamic extended strategy to retrieve the richer syntactic information. Our semantic kernel calculates the protein-protein pair similarity and the context similarity based on two semantic resources: WordNet and Medical Subject Heading (MeSH). We evaluate our method with Support Vector Machine (SVM) and achieve an F-score of 69.40% and an AUC of 92.00%, which show that our method outperforms most of the state-of-the-art systems by integrating semantic information.

Accession Number: WOS:000332845300138

PubMed ID: 24622773

ISSN: 1932-6203

Record 29 of 50

Title: A Topic Clustering Approach to Finding Similar Questions from Large Question and Answer Archives

Author(s): Zhang, WN (Zhang, Wei-Nan); Liu, T (Liu, Ting); Yang, Y (Yang, Yang); Cao, LJ (Cao, Liujuan); Zhang, Y (Zhang, Yu); Ji, RR (Ji, Rongrong)

Source: PLOS ONE **Volume:** 9 **Issue:** 3 **Article Number:** e71511 **DOI:** 10.1371/journal.pone.0071511 **Published:** MAR 4 2014

Abstract: With the blooming of Web 2.0, Community Question Answering (CQA) services such as Yahoo! Answers (http://answers.yahoo.com), WikiAnswer (http://wiki.answers.com), and Baidu Zhidao (http://zhidao.baidu.com), etc., have emerged as alternatives for knowledge and information acquisition. Over time, a large number of question and answer (Q&A) pairs with high quality devoted by human intelligence have been accumulated as a comprehensive knowledge base. Unlike the search engines, which return long lists of results, searching in the CQA services can obtain the correct answers to the question queries by automatically finding similar questions that have already been answered by other users. Hence, it greatly improves the efficiency of the online information retrieval. However, given a question query, finding the similar and well-answered questions is a non-trivial task. The main challenge is the word mismatch between question query (query) and candidate question for retrieval (question). To investigate this problem, in this study, we capture the word semantic similarity between query and question by introducing the topic modeling approach. We then propose an unsupervised machine-learning approach to finding similar questions on CQA Q&A archives. The experimental results show that our proposed approach significantly outperforms the state-of-the-art methods.

Accession Number: WOS:000332475500001

PubMed ID: 24595052

Author Identifiers:

Author	ResearcherID Number	ORCID Number
Cong , Gao	A-3726-2011	

ISSN: 1932-6203

Record 30 of 50

Title: RandomWalks for Knowledge- Based Word Sense Disambiguation

Author(s): Agirre, E (Agirre, Eneko); de Lacalle, OL (Lopez de Lacalle, Oier); Soroa, A (Soroa, Aitor)

Source: COMPUTATIONAL LINGUISTICS **Volume:** 40 **Issue:** 1 **Pages:** 57-84 **DOI:** 10.1162/COLI_a_00164 **Published:** MAR 2014

Abstract: Word Sense Disambiguation (WSD) systems automatically choose the intended meaning of a word in context. In this article we present a WSD algorithm based on random walks over large Lexical Knowledge Bases (LKB). We show that our algorithm performs better than other graph-based methods when run on a graph built from WordNet and eXtended WordNet. Our algorithm and LKB combination compares favorably to other knowledge-based approaches in the literature that use similar knowledge on a variety of English data sets and a data set on Spanish. We include a detailed analysis of the factors that affect the algorithm. The algorithm and the LKBs used are publicly available, and the results easily reproducible.
Accession Number: WOS:000332150100003
Author Identifiers:

Author	ResearcherID Number	ORCID Number
Mendizabal, Elixabete	C-3162-2014	
AGIRRE, ENEKO	H-7323-2015	0000-0003-0775-6057
SOROA ETXABE, AITOR		0000-0001-8573-2654

ISSN: 0891-2017
eISSN: 1530-9312

Record 31 of 50
Title: Comparative Evaluation of Link-Based Approaches for Candidate Ranking in Link-to-Wikipedia Systems
Author(s): Garcia, NF (Fernandez Garcia, Norberto); Fisteus, JA (Arias Fisteus, Jesus); Fernandez, LS (Sanchez Fernandez, Luis)
Source: JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH **Volume:** 49 **Pages:** 733-773 **Published:** 2014
Abstract: In recent years, the task of automatically linking pieces of text (anchors) mentioned in a document to Wikipedia articles that represent the meaning of these anchors has received extensive research attention. Typically, link-to-Wikipedia systems try to find a set of Wikipedia articles that are candidates to represent the meaning of the anchor and, later, rank these candidates to select the most appropriate one. In this ranking process the systems rely on context information obtained from the document where the anchor is mentioned and/or from Wikipedia. In this paper we center our attention in the use of Wikipedia links as context information . In particular , we offer a review of several candidate ranking approaches in the state-of-the-art that rely on Wikipedia link information. In addition , we provide a comparative empirical evaluation of the different approaches on five different corpora: the TAC 2010 corpus and four corpora built from actual Wikipedia articles and news items.
Accession Number: WOS:000335436500001
Author Identifiers:

Author	ResearcherID Number	ORCID Number
Sanchez-Fernandez, Luis	I-3867-2015	0000-0002-9801-4747
Arias Fisteus, Jesus	H-6230-2012	0000-0002-4381-2071

ISSN: 1076-9757
eISSN: 1943-5037

Record 32 of 50
Title: A Grammar-Based Semantic Similarity Algorithm for Natural Language Sentences
Author(s): Lee, MC (Lee, Ming Che); Chang, JW (Chang, Jia Wei); Hsieh, TC (Hsieh, Tung Cheng)
Source: SCIENTIFIC WORLD JOURNAL **Article Number:** 437162 **DOI:** 10.1155/2014/437162 **Published:** 2014
Abstract: This paper presents a grammar and semantic corpus based similarity algorithm for natural language sentences. Natural language, in opposition to "artificial language", such as computer programming languages, is the language used by the general public for daily communication. Traditional information retrieval approaches, such as vector models, LSA, HAL, or even the ontology-based approaches that extend to include concept similarity comparison instead of cooccurrence terms/words, may not always determine the perfect matching while there is no obvious relation or concept overlap between two natural language sentences. This paper proposes a sentence similarity algorithm that takes advantage of corpus-based ontology and grammatical rules to overcome the addressed problems. Experiments on two famous benchmarks demonstrate that the proposed algorithm has a significant performance improvement in sentences/short-texts with arbitrary syntax and structure.
Accession Number: WOS:000334850000001
Author Identifiers:

Author	ResearcherID Number	ORCID Number
Chang, Jia-Wei		0000-0002-4296-4065

ISSN: 1537-744X

Record 33 of 50
Title: Domain Terminology Collection for Semantic Interpretation of Sensor Network Data
Author(s): Hwang, M (Hwang, Myunggwon); Kim, J (Kim, Jinhyung); Gim, J (Gim, Jangwon); Song, SK (Song, Sa-kwang); Jung, H (Jung, Hanmin); Jeong, DH (Jeong, Do-Heon)
Source: INTERNATIONAL JOURNAL OF DISTRIBUTED SENSOR NETWORKS **Article Number:** 827319 **DOI:** 10.1155/2014/827319 **Published:** 2014
Abstract: Many studies have investigated the management of data delivered over sensor networks and attempted to standardize their relations. Sensor data come from numerous tangible and intangible sources, and existing work has focused on the integration and management of the sensor data itself. The data should be interpreted according to the sensor environment and related objects, even though the data type, and even the value, is exactly the same. This means that the sensor data should have semantic connections with all objects, and so a knowledge base that covers all domains should be constructed. In this paper, we suggest a method of domain terminology collection based on Wikipedia category information in order to prepare seed data for such knowledge bases. However, Wikipedia has two weaknesses, namely, loops and unreasonable generalizations in the category structure. To overcome these weaknesses, we utilize a horizontal bootstrapping method for category searches and domain-term collection. Both the category-article and article-link relations defined in Wikipedia are employed as terminology indicators, and we use a new measure to calculate the similarity between categories. By evaluating various aspects of the proposed approach, we show that it outperforms the baseline method, having wider coverage and higher precision. The collected domain terminologies can assist the construction of domain knowledge bases for the semantic interpretation of sensor data.
Accession Number: WOS:000331765400001
Author Identifiers:

Author	ResearcherID Number	ORCID Number
Jung, Hanmin		0000-0001-8690-0664

ISSN: 1550-1329
eISSN: 1550-1477

Record 34 of 50

Title: Cross-Language Opinion Lexicon Extraction Using Mutual-Reinforcement Label Propagation

Author(s): Lin, Z (Lin, Zheng); Tan, SB (Tan, Songbo); Liu, Y (Liu, Yue); Cheng, XQ (Cheng, Xueqi); Xu, XK (Xu, Xueke)

Source: PLOS ONE **Volume:** 8 **Issue:** 11 **Article Number:** e79294 **DOI:** 10.1371/journal.pone.0079294 **Published:** NOV 15 2013

Abstract: There is a growing interest in automatically building opinion lexicon from sources such as product reviews. Most of these methods depend on abundant external resources such as WordNet, which limits the applicability of these methods. Unsupervised or semi-supervised learning provides an optional solution to multilingual opinion lexicon extraction. However, the datasets are imbalanced in different languages. For some languages, the high-quality corpora are scarce or hard to obtain, which limits the research progress. To solve the above problems, we explore a mutual-reinforcement label propagation framework. First, for each language, a label propagation algorithm is applied to a word relation graph, and then a bilingual dictionary is used as a bridge to transfer information between two languages. A key advantage of this model is its ability to make two languages learn from each other and boost each other. The experimental results show that the proposed approach outperforms baseline significantly.

Accession Number: WOS:000327258600029

PubMed ID: 24260190

Author Identifiers:

Author	ResearcherID Number	ORCID Number
Cheng, Xueqi	F-1706-2010	

ISSN: 1932-6203

Record 35 of 50

Title: Metaphor Interpretation Using Paraphrases Extracted from the Web

Author(s): Bollegala, D (Bollegala, Danushka); Shutova, E (Shutova, Ekaterina)

Source: PLOS ONE **Volume:** 8 **Issue:** 9 **Article Number:** e74304 **DOI:** 10.1371/journal.pone.0074304 **Published:** SEP 20 2013

Abstract: Interpreting metaphor is a hard but important problem in natural language processing that has numerous applications. One way to address this task is by finding a paraphrase that can replace the metaphorically used word in a given context. This approach has been previously implemented only within supervised frameworks, relying on manually constructed lexical resources, such as WordNet. In contrast, we present a fully unsupervised metaphor interpretation method that extracts literal paraphrases for metaphorical expressions from the Web. It achieves a precision of 0: 42, which is high for an unsupervised paraphrasing approach. Moreover, the method significantly outperforms both the baseline and the selectional preference-based method of Shutova employed in an unsupervised setting.

Accession Number: WOS:000324768000016

PubMed ID: 24073207

ISSN: 1932-6203

Record 36 of 50

Title: Selectional Preferences for Semantic Role Classification

Author(s): Zapirain, B (Zapirain, Benat); Agirre, E (Agirre, Eneko); Marquez, L (Marquez, Lluis); Surdeanu, M (Surdeanu, Mihai)

Source: COMPUTATIONAL LINGUISTICS **Volume:** 39 **Issue:** 3 **DOI:** 10.1162/COLI_a_00145 **Published:** SEP 2013

Abstract: This paper focuses on a well-known open issue in Semantic Role Classification (SRC) research: the limited influence and sparseness of lexical features. We mitigate this problem using models that integrate automatically learned selectional preferences (SP). We explore a range of models based on WordNet and distributional-similarity SPs. Furthermore, we demonstrate that the SRC task is better modeled by SP models centered on both verbs and prepositions, rather than verbs alone. Our experiments with SP-based models in isolation indicate that they outperform a lexical baseline with 20 F-1 points in domain and almost 40 F-1 points out of domain. Furthermore, we show that a state-of-the-art SRC system extended with features based on selectional preferences performs significantly better, both in domain (17% error reduction) and out of domain (13% error reduction). Finally, we show that in an end-to-end semantic role labeling system we obtain small but statistically significant improvements, even though our modified SRC model affects only approximately 4% of the argument candidates. Our post hoc error analysis indicates that the SP-based features help mostly in situations where syntactic information is either incorrect or insufficient to disambiguate the correct role.

Accession Number: WOS:000325864800006

Author Identifiers:

Author	ResearcherID Number	ORCID Number
Mendizabal, Elixabete	C-3162-2014	
AGIRRE, ENEKO	H-7323-2015	0000-0003-0775-6057

ISSN: 0891-2017

eISSN: 1530-9312

Record 37 of 50

Title: User Evaluation of the Effects of a Text Simplification Algorithm Using Term Familiarity on Perception, Understanding, Learning, and Information Retention

Author(s): Leroy, G (Leroy, Gondy); Endicott, JE (Endicott, James E.); Kauchak, D (Kauchak, David); Mouradi, O (Mouradi, Obay); Just, M (Just, Melissa)

Source: JOURNAL OF MEDICAL INTERNET RESEARCH **Volume:** 15 **Issue:** 7 **Pages:** 191-203 **Article Number:** UNSP e144 **DOI:** 10.2196/jmir.2569 **Published:** JUL 2013

Abstract: Background: Adequate health literacy is important for people to maintain good health and manage diseases and injuries. Educational text, either retrieved from the Internet or provided by a doctor's office, is a popular method to communicate health-related information. Unfortunately, it is difficult to write text that is easy to understand, and existing approaches, mostly the application of readability formulas, have not convincingly been shown to reduce the difficulty of text.

Objective: To develop an evidence-based writer support tool to improve perceived and actual text difficulty. To this end, we are developing and testing algorithms that automatically identify difficult sections in text and provide appropriate, easier alternatives; algorithms that effectively reduce text difficulty will be included in the support tool. This work describes the user evaluation with an independent writer of an automated simplification algorithm using term familiarity.

Methods: Term familiarity indicates how easy words are for readers and is estimated using term frequencies in the Google Web Corpus. Unfamiliar words are algorithmically identified and tagged for potential replacement. Easier alternatives consisting of synonyms, hypernyms, definitions, and semantic types are extracted from WordNet, the Unified Medical Language System (UMLS), and Wiktionary and ranked for a writer to choose from to simplify the text. We conducted a controlled user study with a representative writer who used our simplification algorithm to simplify texts. We tested the impact with representative consumers. The key independent variable of our study is lexical simplification, and we measured its effect on both perceived and actual text difficulty. Participants were recruited from Amazon's Mechanical Turk website. Perceived difficulty was measured with 1 metric, a 5-point Likert scale. Actual difficulty was measured with 3 metrics: 5 multiple-choice questions alongside each text to measure understanding, 7 multiple-choice questions without the text for learning, and 2 free recall questions for information retention.

Results: Ninety-nine participants completed the study. We found strong beneficial effects on both perceived and actual difficulty. After simplification, the text was perceived as simpler (P<.001) with simplified text scoring 2.3 and original text 3.2 on the 5-point Likert scale (score 1: easiest). It also led to better understanding of the text (P<.001) with 11% more correct answers with simplified text (63% correct) compared to the original (52% correct). There was more learning with 18% more correct answers after reading simplified text compared to 9% more correct answers after reading the original text (P=.003). There was no significant effect on free recall.

Conclusions: Term familiarity is a valuable feature in simplifying text. Although the topic of the text influences the effect size, the results were convincing and consistent.

Accession Number: WOS:000324620300015
PubMed ID: 23903235
Author Identifiers:

Author	ResearcherID Number	ORCID Number
Endicott, James	K-1312-2012	0000-0002-8434-0859

ISSN: 1438-8871

Record 38 of 50
Title: Structural Similarities between Brain and Linguistic Data Provide Evidence of Semantic Relations in the Brain
Author(s): Crangle, CE (Crangle, Colleen E.); Perreau-Guimaraes, M (Perreau-Guimaraes, Marcos); Suppes, P (Suppes, Patrick)
Source: PLOS ONE **Volume:** 8 **Issue:** 6 **Article Number:** e65366 **DOI:** 10.1371/journal.pone.0065366 **Published:** JUN 14 2013
Abstract: This paper presents a new method of analysis by which structural similarities between brain data and linguistic data can be assessed at the semantic level. It shows how to measure the strength of these structural similarities and so determine the relatively better fit of the brain data with one semantic model over another. The first model is derived from WordNet, a lexical database of English compiled by language experts. The second is given by the corpus-based statistical technique of latent semantic analysis (LSA), which detects relations between words that are latent or hidden in text. The brain data are drawn from experiments in which statements about the geography of Europe were presented auditorily to participants who were asked to determine their truth or falsity while electroencephalographic (EEG) recordings were made. The theoretical framework for the analysis of the brain and semantic data derives from axiomatizations of theories such as the theory of differences in utility preference. Using brain-data samples from individual trials time-locked to the presentation of each word, ordinal relations of similarity differences are computed for the brain data and for the linguistic data. In each case those relations that are invariant with respect to the brain and linguistic data, and are correlated with sufficient statistical strength, amount to structural similarities between the brain and linguistic data. Results show that many more statistically significant structural similarities can be found between the brain data and the WordNet-derived data than the LSA-derived data. The work reported here is placed within the context of other recent studies of semantics and the brain. The main contribution of this paper is the new method it presents for the study of semantics and the brain and the focus it permits on networks of relations detected in brain data and represented by a semantic model.
Accession Number: WOS:000320363300016
PubMed ID: 23799009
ISSN: 1932-6203

Record 39 of 50
Title: Statistical Metaphor Processing
Author(s): Shutova, E (Shutova, Ekaterina); Teufel, S (Teufel, Simone); Korhonen, A (Korhonen, Anna)
Source: COMPUTATIONAL LINGUISTICS **Volume:** 39 **Issue:** 2 **Pages:** 301-353 **DOI:** 10.1162/COLI_a_00124 **Published:** JUN 2013
Abstract: Metaphor is highly frequent in language, which makes its computational processing indispensable for real-world NLP applications addressing semantic tasks. Previous approaches to metaphor modeling rely on task-specific hand-coded knowledge and operate on a limited domain or a subset of phenomena. We present the first integrated open-domain statistical model of metaphor processing in unrestricted text. Our method first identifies metaphorical expressions in running text and then paraphrases them with their literal paraphrases. Such a text-to-text model of metaphor interpretation is compatible with other NLP applications that can benefit from metaphor resolution. Our approach is minimally supervised, relies on the state-of-the-art parsing and lexical acquisition technologies (distributional clustering and selectional preference induction), and operates with a high accuracy.
Accession Number: WOS:000318556900003
ISSN: 0891-2017
eISSN: 1530-9312

Record 40 of 50
Title: A Bottom-Up Approach for Automatically Grouping Sensor Data Layers by their Observed Property
Author(s): Knoechel, B (Knoechel, Ben); Huang, CY (Huang, Chih-Yuan); Liang, SHL (Liang, Steve H. L.)
Source: ISPRS INTERNATIONAL JOURNAL OF GEO-INFORMATION **Volume:** 2 **Issue:** 1 **Pages:** 1-26 **DOI:** 10.3390/ijgi2010001 **Published:** MAR 2013
Abstract: The Sensor Web is a growing phenomenon where an increasing number of sensors are collecting data in the physical world, to be made available over the Internet. To help realize the Sensor Web, the Open Geospatial Consortium (OGC) has developed open standards to standardize the communication protocols for sharing sensor data. Spatial Data Infrastructures (SDIs) are systems that have been developed to access, process, and visualize geospatial data from heterogeneous sources, and SDIs can be designed specifically for the Sensor Web. However, there are problems with interoperability associated with a lack of standardized naming, even with data collected using the same open standard. The objective of this research is to automatically group similar sensor data layers. We propose a methodology to automatically group similar sensor data layers based on the phenomenon they measure. Our methodology is based on a unique bottom-up approach that uses text processing, approximate string matching, and semantic string matching of data layers. We use WordNet as a lexical database to compute word pair similarities and derive a set-based dissimilarity function using those scores. Two approaches are taken to group data layers: mapping is defined between all the data layers, and clustering is performed to group similar data layers. We evaluate the results of our methodology.
Accession Number: WOS:000209465600001
ISSN: 2220-9964

Record 41 of 50
Title: Learning to Predict from Textual Data
Author(s): Radinsky, K (Radinsky, Kira); Davidovich, S (Davidovich, Sagie); Markovitch, S (Markovitch, Shaul)
Source: JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH **Volume:** 45 **Pages:** 641-684 **Published:** 2012
Abstract: Given a current news event, we tackle the problem of generating plausible predictions of future events it might cause. We present a new methodology for modeling and predicting such future news events using machine learning and data mining techniques. Our Pundit algorithm generalizes examples of causality pairs to infer a causality predictor. To obtain precisely labeled causality examples, we mine 150 years of news articles and apply semantic natural language modeling techniques to headlines containing certain predefined causality patterns. For generalization, the model uses a vast number of world knowledge ontologies. Empirical evaluation on real news articles shows that our Pundit algorithm performs as well as non-expert humans.
Accession Number: WOS:000312767500001
ISSN: 1076-9757
eISSN: 1943-5037

Record 42 of 50
Title: Semantic Similarity Measures Applied to an Ontology for Human-Like Interaction
Author(s): Albacete, E (Albacete, Esperanza); Calle, J (Calle, Javier); Castro, E (Castro, Elena); Cuadra, D (Cuadra, Dolores)
Source: JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH **Volume:** 44 **Pages:** 397-421 **Published:** 2012
Abstract: The focus of this paper is the calculation of similarity between two concepts from an ontology for a Human-Like Interaction system. In order to facilitate this calculation, a similarity function is proposed based on five dimensions (sort, compositional, essential, restrictive and descriptive) constituting the structure of ontological knowledge. The paper includes a proposal for computing a similarity function for each dimension of knowledge. Later on, the similarity values obtained are weighted and aggregated to obtain a global similarity measure. In order to calculate those weights associated to each dimension, four training methods have been

proposed. The training methods differ in the element to fit: the user, concepts or pairs of concepts, and a hybrid approach. For evaluating the proposal, the knowledge base was fed from WordNet and extended by using a knowledge editing toolkit (Cognos). The evaluation of the proposal is carried out through the comparison of system responses with those given by human test subjects, both providing a measure of the soundness of the procedure and revealing ways in which the proposal may be improved.

Accession Number: WOS:000305881400001

ISSN: 1076-9757

Record 43 of 50

Title: STRUCTURAL ANALYSIS OF THE CURRENT ESTONIAN WORDNET

Author(s): Lohk, A (Lohk, Ahti); Vohandu, L (Vohandu, Leo)

Source: EESTI RAKENDUSLINGVISTIKA UHINGU AASTARAAMAT **Volume:** 8 **Pages:** 139-151 **DOI:** 10.5128/ERYa8.09 **Published:** 2012

Abstract: Control of any expanding and developing system requires a feedback control mechanism to evaluate the normal trends of the system and also the unsystematic steps. Experience has shown that often more progress is achieved in a field if non-specialists intervene. For this reason the authors (from the field of informatics) suggest special data processing and visualization methods for the Estonian Wordnet. The images presented are suitable as a feedback mechanism for every Wordnet-style linked system, for semantic analysis and examination of regularities. They allow the lexicographer to use an innovative approach to study the hidden structures of the system and make corrections if needed. Several complex data analysis methods have been used (bipartite graph clustering, minimization of interval graph crossing numbers) and the results presented in this article. Positive feedback from the makers and maintainers of the Estonian Wordnet allows us to be sure that the methods presented would be useful for other language Wordnets also.

Accession Number: WOS:000303361300009

ISSN: 1736-2563

Record 44 of 50

Title: The CQC Algorithm: Cycling in Graphs to Semantically Enrich and Enhance a Bilingual Dictionary

Author(s): Flati, T (Flati, Tiziano); Navigli, R (Navigli, Roberto)

Source: JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH **Volume:** 43 **Pages:** 135-171 **Published:** 2012

Abstract: Bilingual machine-readable dictionaries are knowledge resources useful in many automatic tasks. However, compared to monolingual computational lexicons like WordNet, bilingual dictionaries typically provide a lower amount of structured information such as lexical and semantic relations, and often do not cover the entire range of possible translations for a word of interest. In this paper we present Cycles and Quasi-Cycles (CQC), a novel algorithm for the automated disambiguation of ambiguous translations in the lexical entries of a bilingual machine-readable dictionary. The dictionary is represented as a graph, and cyclic patterns are sought in this graph to assign an appropriate sense tag to each translation in a lexical entry. Further, we use the algorithm's output to improve the quality of the dictionary itself, by suggesting accurate solutions to structural problems such as misalignments, partial alignments and missing entries. Finally, we successfully apply CQC to the task of synonym extraction.

Accession Number: WOS:000300418600001

ISSN: 1076-9757

Record 45 of 50

Title: A WordNet-Based Near-Synonyms and Similar-Looking Word Learning System

Author(s): Sun, KT (Sun, Koun-Tem); Huang, YM (Huang, Yueh-Min); Liu, MC (Liu, Ming-Chi)

Source: EDUCATIONAL TECHNOLOGY & SOCIETY **Volume:** 14 **Issue:** 1 **Pages:** 121-134 **Published:** JAN 2011

Abstract: Near-Synonyms and Similar-Looking (NSSL) words can create confusion for English as Foreign Language Learners as a result of a type of lexical error that often occurs when they confuse similar-looking words that are near synonyms to have the same meaning. Particularly, this may occur if the similar-looking words have the same translated meaning. This study proposes a method to find these NSSL words and designed three experiments to investigate whether NSSL matching exercises could increase Chinese EFL learners' awareness of NSSL words. Three primary findings arose from the study. First, a performance evaluation of the experiment showed good results and determined that the method extracted suitable NSSL words whose meaning EFL learners may confuse. Secondly, the analysis results of the evaluation of Computer Assisted Language Learning (CALL) software showed that this system is practical for language learning, but lacks authenticity. Thirdly, a total of ninety-two Chinese students participated in this study and the findings indicated that students increased awareness of NSSL words and improved in ability of NSSL word distinction while still maintaining the knowledge one month after they had completed the matching exercises. Additionally, students' feedback expressed that they had benefited from discovery learning and that they thought it was not difficult to discover the differences among NSSL words. Further research might extend the method proposed in this study to distracter choice of automatic question generation.

Accession Number: WOS:000287796900011

ISSN: 1436-4522

Record 46 of 50

Title: Extraction of Method Signatures from Ontology Towards Reusability for the Given System Requirement Specification

Author(s): Sagayaraj, S (Sagayaraj, S.); Ganapathy, G (Ganapathy, Gopinath)

Edited by: Ao SI; Gelman L; Hukins DWL; Hunter A; Korsunsky AM

Source: WORLD CONGRESS ON ENGINEERING, WCE 2011, VOL III **Book Series:** Lecture Notes in Engineering and Computer Science **Pages:** 1877-1882 **Published:** 2011

Abstract: Software reuse improves productivity, quality, and maintainability of software products. Only few completed projects are achieved and documented. The method signatures in a completed project are stored in the Ontology and the source code components are stored in Hadoop Distributed File System (HDFS). Methods are needed for the new project can be extracted from the Ontology using Software Requirement Specification (SRS) document. UML design document will evolve after many phases from SRS and hence this work proposes a new framework to extract keywords from SRS and estimate the number of new methods to be developed and count the number of methods that can be reused from the Ontology. The SRS document for the project consists of purpose, scope, system requirements, functional requirements and non-functional requirements as metadata. The SRS document is given as input and the keywords are extracted. The keywords are searched in Ontology for the similar method prototypes and the appropriate code components would be extracted from the HDFS. These methods are integrated in the new project with a review process. The implementation is provided with the sample SRS text. The keywords are extracted and matched with the Ontology. The reusability is measured using reuse metrics, quality, and knowledge growth.

Accession Number: WOS:000393014000013

Conference Title: World Congress on Engineering (WCE 2011)

Conference Date: JUL 06-08, 2011

Conference Location: Imperial Coll, London, UNITED KINGDOM

Conference Sponsors: Int Assoc Engineers, IAENG, Soc Artificial Intelligence, IAENG, Soc Bioinformat, IAENG, Soc Computer Sci, IAENG, Soc Data Min, IAENG, Soc Elect Engn, IAENG, Soc Imagl Engn, IAENG, Soc Ind Engn, IAENG, Soc Informat Syst Engn, IAENG, Soc Internet Comput & Web Serv, IAENG, Soc Mech Engn, IAENG, Soc Operat Res, IAENG, Soc Sci Comput, IAENG, Soc Software Engn, IAENG, Soc Wireless Engn

Conference Host: Imperial Coll

ISSN: 2078-0958

ISBN: 978-988-19251-5-2

Record 47 of 50

Title: A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences

Author(s): Erk, K (Erk, Katrin); Pado, S (Pado, Sebastian); Pado, U (Pado, Ulrike)

Source: COMPUTATIONAL LINGUISTICS **Volume:** 36 **Issue:** 4 **Pages:** 723-763 **DOI:** 10.1162/coli_a_00017 **Published:** DEC 2010

Abstract: We present a vector space-based model for selectional preferences that predicts plausibility scores for argument headwords. It does not require any lexical resources (such as WordNet). It can be trained either on one corpus with syntactic annotation, or on a combination of a small semantically annotated primary corpus and a large, syntactically analyzed generalization corpus. Our model is able to predict inverse selectional preferences, that is, plausibility scores for predicates given argument heads.
We evaluate our model on one NLP task (pseudo-disambiguation) and one cognitive task (prediction of human plausibility judgments), gauging the influence of different parameters and comparing our model against other model classes. We obtain consistent benefits from using the disambiguation and semantic role information provided by a semantically tagged primary corpus. As for parameters, we identify settings that yield good performance across a range of experimental conditions. However, frequency remains a major influence of prediction quality, and we also identify more robust parameter settings suitable for applications with many infrequent items.

Accession Number: WOS:000285382400007

Author Identifiers:

Author	ResearcherID Number	ORCID Number
Pado, Sebastian	F-4883-2016	0000-0002-7529-6825

ISSN: 0891-2017

Record 48 of 50

Title: Principal Semantic Components of Language and the Measurement of Meaning

Author(s): Samsonovic, AV (Samsonovic, Alexei V.); Ascoli, GA (Ascoli, Giorgio A.)

Source: PLOS ONE **Volume:** 5 **Issue:** 6 **Article Number:** e10921 **DOI:** 10.1371/journal.pone.0010921 **Published:** JUN 11 2010

Abstract: Metric systems for semantics, or semantic cognitive maps, are allocations of words or other representations in a metric space based on their meaning. Existing methods for semantic mapping, such as Latent Semantic Analysis and Latent Dirichlet Allocation, are based on paradigms involving dissimilarity metrics. They typically do not take into account relations of antonymy and yield a large number of domain-specific semantic dimensions. Here, using a novel self-organization approach, we construct a low-dimensional, context-independent semantic map of natural language that represents simultaneously synonymy and antonymy. Emergent semantics of the map principal components are clearly identifiable: the first three correspond to the meanings of "good/bad" (valence), "calm/excited" (arousal), and "open/closed" (freedom), respectively. The semantic map is sufficiently robust to allow the automated extraction of synonyms and antonyms not originally in the dictionaries used to construct the map and to predict connotation from their coordinates. The map geometric characteristics include a limited number (similar to 4) of statistically significant dimensions, a bimodal distribution of the first component, increasing kurtosis of subsequent (unimodal) components, and a U-shaped maximum-spread planar projection. Both the semantic content and the main geometric features of the map are consistent between dictionaries (Microsoft Word and Princeton's WordNet), among Western languages (English, French, German, and Spanish), and with previously established psychometric measures. By defining the semantics of its dimensions, the constructed map provides a foundational metric system for the quantitative analysis of word meaning. Language can be viewed as a cumulative product of human experiences. Therefore, the extracted principal semantic dimensions may be useful to characterize the general semantic dimensions of the content of mental states. This is a fundamental step toward a universal metric system for semantics of human experiences, which is necessary for developing a rigorous science of the mind.

Accession Number: WOS:000278662900002

PubMed ID: 20552009

ISSN: 1932-6203

Record 49 of 50

Title: The Noisy Channel Mode for Unsupervised Word Sense Disambiguation

Author(s): Yuret, D (Yuret, Deniz); Yatbaz, MA (Yatbaz, Mehmet Ali)

Source: COMPUTATIONAL LINGUISTICS **Volume:** 36 **Issue:** 1 **Pages:** 111-127 **DOI:** 10.1162/coli.2010.36.1.36103 **Published:** MAR 2010

Abstract: We introduce a generative probabilistic model, the noisy channel model, for unsupervised word sense disambiguation. In our model, each context C is modeled as a distinct channel through which the speaker intends to transmit a particular meaning S using a possibly ambiguous word W. To reconstruct the intended meaning the hearer uses the distribution of possible meanings in the given context P(S|C) and possible words that can express each meaning P(W|S). We assume P(W|S) is independent of the context and estimate it using WordNet sense frequencies. The main problem of unsupervised WSD is estimating context-dependent P(S|C) without access to any sense-tagged text. We show one way to solve this problem using a statistical language model based on large amounts of untagged text. Our model uses coarse-grained semantic classes for S internally and we explore the effect of using different levels of granularity on WSD performance. The system outputs fine-grained senses for evaluation, and its performance on noun disambiguation is better than most previously reported unsupervised systems and close to the best supervised systems.

Accession Number: WOS:000275310400004

ISSN: 0891-2017

Record 50 of 50

Title: Interaction Chain Patterns of Online Text Construction with Lexical Cohesion

Author(s): Yeh, HC (Yeh, Hui-Chin); Yang, YF (Yang, Yu-Fen); Wong, WK (Wong, Wing-Kwong)

Source: EDUCATIONAL TECHNOLOGY & SOCIETY **Volume:** 13 **Issue:** 1 **Pages:** 55-68 **Published:** JAN 2010

Abstract: This study aims at arousing college students' metacognition in detecting lexical cohesion during online text construction as WordNet served as a lexical resource. A total of 83 students were requested to construct texts through sequences of actions identified as interaction chains in this study. Interaction chains are grouped and categorized as a meaningful entity in order to investigate the students' thinking process and behavior in general and to understand the interaction between the computer and the students in particular. From the interaction chains, it was found that some students revised incorrect sentences to correct ones. In making correct revision, they needed to assess incoming information, interpret and organize textual information, engage in thinking what they know, monitor their own meaning construction process, and take remedial actions to reach comprehension. The rate of correct sentence selection increased from 34.04% to 55.02% in three sequential text construction tasks. The recognition of lexical cohesion was found to be a determining factor for successful construction of a text.

Accession Number: WOS:000274865500006

Author Identifiers:

Author	ResearcherID Number	ORCID Number
Yeh, Hui-Chin	O-1256-2013	

ISSN: 1436-4522

Close

Web of Science™
Page 1 (Records 1 -- 50)

[1]

Print