Close

**Web of Science™**
**Page 1 (Records 1 -- 15)**
[ 1 ]

Print

**Record 1 of 15**
**Title:** Exploiting Semantic Role Resources for Preposition Disambiguation
**Author(s):** O'Hara, T (O'Hara, Tom); Wiebe, J (Wiebe, Janyce)
**Source:** COMPUTATIONAL LINGUISTICS **Volume:** 35 **Issue:** 2 **Pages:** 151-184 **DOI:** 10.1162/coli.06-79-prep15 **Published:** JUN 2009
**Abstract:** This article describes how semantic role resources can be exploited for preposition disambiguation. The main resources include the semantic role annotations provided by the Penn Treebank and FrameNet tagged corpora. The resources also include the assertions contained in the Factotum knowledge base, as well as information from Cyc and Conceptual Graphs. A common inventory is derived from these in support of definition analysis, which is the motivation for this work.
The disambiguation concentrates on relations indicated by prepositional phrases, and is framed as word-sense disambiguation for the preposition in question. A new type of feature for word-sense disambiguation is introduced, using WordNet hypernyms as collocations rather than just words. Various experiments over the Penn Treebank and FrameNet data are presented, including prepositions classified separately versus together, and illustrating the effects of filtering. Similar experimentation is done over the Factotum data, including a method for inferring likely preposition usage from corpora, as knowledge bases do not generally indicate how relationships are expressed in English ( in contrast to the explicit annotations on this in the Penn Treebank and FrameNet). Other experiments are included with the FrameNet data mapped into the common relation inventory developed for definition analysis, illustrating how preposition disambiguation might be applied in lexical acquisition.
**Accession Number:** WOS:000272350800002
**ISSN:** 0891-2017

**Record 2 of 15**
**Title:** Wikipedia-based Semantic Interpretation for Natural Language Processing
**Author(s):** Gabrilovich, E (Gabrilovich, Evgeniy); Markovitch, S (Markovitch, Shaul)
**Source:** JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH **Volume:** 34 **Pages:** 443-498 **Published:** 2009
**Abstract:** Adequate representation of natural language semantics requires access to vast amounts of common sense and domain-specific world knowledge. Prior work in the field was based on purely statistical techniques that did not make use of background knowledge, on limited lexicographic knowledge bases such as Word Net, or on huge manual efforts such as the CYC project. Here we propose novel method, called Explicit Semantic Analysis (ESA), for fine-grained semantic interpretation of unrestricted natural language texts. Our method represents meaning in a high-dimensional space of concepts derived from Wikipedia, the largest encyclopedia in existence. We explicitly represent the meaning of any text in terms of Wikipedia-based concepts. We evaluate the effectiveness of our method on text categorization and on computing the degree of semantic relatedness between fragments of natural language text. Using ESA results insignificant improvements over the previous state of the art in both tasks. Importantly, due to the use of natural concepts, the ESA model is easy to explain to human users.
**Accession Number:** WOS:000264721000003
**ISSN:** 1076-9757
**eISSN:** 1943-5037

**Record 3 of 15**
**Title:** Discovering gene annotations in biomedical text databases
**Author(s):** Cakmak, A (Cakmak, Ali); Ozsoyoglu, G (Ozsoyoglu, Gultekin)
**Source:** BMC BIOINFORMATICS **Volume:** 9 **Article Number:** 143 **DOI:** 10.1186/1471-2105-9-143 **Published:** MAR 6 2008
**Abstract:** Background: Genes and gene products are frequently annotated with Gene Ontology concepts based on the evidence provided in genomics articles. Manually locating and curating information about a genomic entity from the biomedical literature requires vast amounts of human effort. Hence, there is clearly a need forautomated computational tools to annotate the genes and gene products with Gene Ontology concepts by computationally capturing the related knowledge embedded in textual data.
Results: In this article, we present an automated genomic entity annotation system, GEANN, which extracts information about the characteristics of genes and gene products in article abstracts from PubMed, and translates the discoveredknowledge into Gene Ontology ( GO) concepts, a widely-used standardized vocabulary of genomic traits. GEANN utilizes textual "extraction patterns", and a semantic matching framework to locate phrases matching to a pattern and produce Gene Ontology annotations for genes and gene products.
In our experiments, GEANN has reached to the precision level of 78% at therecall level of 61%. On a select set of Gene Ontology concepts, GEANN either outperforms or is comparable to two other automated annotation studies. Use of WordNet for semantic pattern matching improves the precision and recall by 24% and 15%, respectively, and the improvement due to semantic pattern matching becomes more apparent as the Gene Ontology terms become more general.
Conclusion: GEANN is useful for two distinct purposes: (i) automating the annotation of genomic entities with Gene Ontology concepts, and (ii) providing existing annotations with additional "evidence articles" from the literature. The use of textual extraction patterns that are constructed based on the existing annotations achieve high precision. The semantic pattern matching framework provides a more flexible pattern matching scheme with respect to "exactmatching" with the advantage of locating approximate pattern occurrences with similar semantics. Relatively low recall performance of our pattern-based approach may be enhanced either by employing a probabilistic annotation framework based on the annotation neighbourhoods in textual data, or, alternatively, the statistical enrichment threshold may be adjusted to lower values for applications that put more value on achieving higher recall values.
**Accession Number:** WOS:000255283100002
**PubMed ID:** 18325104
**ISSN:** 1471-2105

**Record 4 of 15**
**Title:** HOW DO WE COMPREHEND THE CONCEPTS?
**Author(s):** Vainik, E (Vainik, Ene); Kirt, T (Kirt, Toomas)
**Source:** EESTI RAKENDUSLINGVISTIKA UHINGU AASTARAAMAT **Volume:** 4 **Pages:** 225-245 **Published:** 2008
**Abstract:** Scholars differ over the meaning of concept. Having first belonged to the field of logic and philosophy, concepts are now of interest also for (neuro)linguists, psychologists, cognitive scientists and and specialists in informatics. The aim of the article is to provide an overview of different theoretical approaches to concept and to demonstrate the relations of those approaches to certain tendencies in linguistic semantics.
In the classical approach it is held important that concepts should be defined in terms of necessary and sufficient conditions. In linguistics the approach based on very precise differentia has become the core of semantic component analysis, while its application has been central, e.g., in the field of terminology. Prototype theory, largely drawing on psychologial reliability, explains concepts as radial categories organized around the most representative members or category prototypes. For linguistics this approach implied that identification of the meanings of a word requires consideration of its typical as well as less typical uses alongside their different statuses. The theory theory aims at explaining the essence of concepts by means of certain more general knowledge structures (theories) that the concepts are part of. This approach has been used a lot for different semantic representations and description models of semantic content, such as frames, scripts or scenarios.
There have also been attempts to describe concepts in spatial terms, using dimensional representations. Cognitive linguistics, as well as cognitive sciences in general, are rather prone to construct some kind of semantic or conceptual spaces. Another type of concept explication is a network model of knowledge, which serves as basis for the so-called encyclopaedic approach to semantics. Here a concept is described by its relations with other concepts in the network of knowledge. The main linguistic applications are thesaurus dictionaries and electronic resources like WordNet. The dynamic approach views concepts as individual and variable situation-related processes going on in human brain, which have to do with processing and storing sensory data. It has been found that basically same parts of the neural

substrate are involved in representing concepts (e. g. such as is represented by the verb grasp) as well as their underlying events (some act of grasping). Concepts are said to be simulations or reenactments of sensorimotor representations of events. That approach lays the basis for the so-called interactive semantics, arguing that human (incl. bodily) interaction with the external world is responsible for the development of the embodied concepts or schemes. Evolutionary approach sets out to reveal the mechanism of symbol grounding and concept formation, trying to explain it by ubiquitous loops of positive and negative feedback. It is vital for an organism, on the one hand, to tell hazardous situations from beneficial ones, and on the other hand, to assess what are the affordances of the situations.
Instead of aiming at the one and only theory of what the concepts really are the article concludes that different theoretical approaches focus on different aspects of concepts. Description, explication and modelling of linguistic concepts are three different objectives, each thus preferring a different semantic representation.
**Accession Number:** WOS:000260404800014
**ISSN:** 1736-2563

---

**Record 5 of 15**
**Title:** On the use of automatically acquired examples for all-nouns Word Sense Disambiguation
**Author(s):** Martinez, D (Martinez, David); de Lacalle, OL (de Lacalle, Oier Lopez); Agirre, E (Agirre, Eneko)
**Source:** JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH **Volume:** 33 **Pages:** 79-107 **Published:** 2008
**Abstract:** This article focuses on Word Sense Disambiguation (WSD), which is a Natural Language Processing task that is thought to be important for many Language Technology applications, such as Information Retrieval, Information Extraction, or Machine Translation. One of the main issues preventing the deployment of WSD technology is the lack of training examples for Machine Learning systems, also known as the Knowledge Acquisition Bottleneck. A method which has been shown to work for small samples of words is the automatic acquisition of examples. We have previously shown that one of the most promising example acquisition methods scales up and produces a freely available database of 150 million examples from Web snippets for all polysemous nouns in WordNet. This paper focuses on the issues that arise when using those examples, all alone or in addition to manually tagged examples, to train a supervised WSD system for all nouns. The extensive evaluation on both lexical-sample and all-words Senseval benchmarks shows that we are able to improve over commonly used baselines and to achieve top-rank performance. The good use of the prior distributions from the senses proved to be a crucial factor.
**Accession Number:** WOS:000259611900002
**Author Identifiers:**

| Author | ResearcherID Number | ORCID Number |
|---|---|---|
| Mendizabal, Elixabete | C-3162-2014 | |
| AGIRRE, ENEKO | H-7323-2015 | 0000-0003-0775-6057 |
| Martinez, David | | 0000-0002-8969-9318 |

**ISSN:** 1076-9757
**eISSN:** 1943-5037

---

**Record 6 of 15**
**Title:** A UMLS-based spell checker for natural language processing in vaccine safety
**Author(s):** Tolentino, HD (Tolentino, Herman D.); Matters, MD (Matters, Michael D.); Walop, W (Walop, Wikke); Law, B (Law, Barbara); Tong, W (Tong, Wesley); Liu, F (Liu, Fang); Fontelo, P (Fontelo, Paul); Kohl, K (Kohl, Katrin); Payne, DC (Payne, Daniel C.)
**Source:** BMC MEDICAL INFORMATICS AND DECISION MAKING **Volume:** 7 **Article Number:** 3 **DOI:** 10.1186/1472-6947-7-3 **Published:** FEB 12 2007
**Abstract:** Background: The Institute of Medicine has identified patient safety as a key goal for health care in the United States. Detecting vaccine adverse events is an important public health activity that contributes to patient safety. Reports about adverse events following immunization (AEFI) from surveillance systems contain free-text components that can be analyzed using natural language processing. To extract Unified Medical Language System ( UMLS) concepts from free text and classify AEFI reports based on concepts they contain, we first needed to clean the text by expanding abbreviations and shortcuts and correcting spelling errors. Our objective in this paper was to create a UMLS-based spelling error correction tool as a first step in the natural language processing (NLP) pipeline for AEFI reports.
Methods: We developed spell checking algorithms using open source tools. We used de-identified AEFI surveillance reports to create free-text data sets for analysis. After expansion of abbreviated clinical terms and shortcuts, we performed spelling correction in four steps: ( 1) error detection, ( 2) word list generation, ( 3) word list disambiguation and ( 4) error correction. We then measured the performance of the resulting spell checker by comparing it to manual correction.
Results: We used 12,056 words to train the spell checker and tested its performance on 8,131 words. During testing, sensitivity, specificity, and positive predictive value (PPV) for the spell checker were 74% ( 95% CI: 74 - 75), 100% ( 95% CI: 100 - 100), and 47% ( 95% CI: 46% - 48%), respectively.
Conclusion: We created a prototype spell checker that can be used to process AEFI reports. We used the UMLS Specialist Lexicon as the primary source of dictionary terms and the WordNet lexicon as a secondary source. We used the UMLS as a domain-specific source of dictionary terms to compare potentially misspelled words in the corpus. The prototype sensitivity was comparable to currently available tools, but the specificity was much superior. The slow processing speed may be improved by trimming it down to the most useful component algorithms. Other investigators may find the methods we developed useful for cleaning text using lexicons specific to their area of interest.
**Accession Number:** WOS:000246516100001
**PubMed ID:** 17295907
**ISSN:** 1472-6947

---

**Record 7 of 15**
**Title:** Knowledge derived from Wikipedia for computing semantic relatedness
**Author(s):** Ponzetto, SP (Ponzetto, Simone Paolo); Strube, M (Strube, Michael)
**Source:** JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH **Volume:** 30 **Pages:** 181-212 **Published:** 2007
**Abstract:** Wikipedia provides a semantic network for computing semantic relatedness in a more structured fashion than a search engine and with more coverage than WordNet. We present experiments on using Wikipedia for computing semantic relatedness and compare it to WordNet on various bench-marking datasets. Existing relatedness measures perform better using Wikipedia than a baseline given by Google counts, and we show that Wikipedia outperforms WordNet on some datasets. We also address the question whether and how Wikipedia can be integrated into NLP applications as a knowledge base. Including Wikipedia improves the performance of a machine learning based coreference resolution system, indicating that it represents a valuable resource for NLP applications. Finally, we show that our method can be easily used for languages other than English by computing semantic relatedness for a German dataset.
**Accession Number:** WOS:000250051500003
**ISSN:** 1076-9757

---

**Record 8 of 15**
**Title:** NP animacy identification for anaphora resolution
**Author(s):** Orasan, C (Orasan, Constantin); Evans, R (Evans, Richard)
**Source:** JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH **Volume:** 29 **Pages:** 79-103 **Published:** 2007
**Abstract:** In anaphora resolution for English, animacy identification can play an integral role in the application of agreement restrictions between pronouns and candidates, and as a result, can improve the accuracy of anaphora resolution systems. In this paper, two methods for animacy identification are proposed and evaluated using intrinsic and extrinsic measures. The first method is a rule-based one which uses information about the unique beginners in WordNet to classify NPs on the basis of their animacy. The second method relies on a machine learning algorithm which exploits a WordNet enriched with animacy information for each sense. The effect of word sense disambiguation on the two methods is also assessed. The intrinsic evaluation reveals that the machine learning method reaches human levels of performance. The extrinsic evaluation demonstrates that animacy identification can be beneficial in anaphora resolution, especially in the cases where

animate entities are identified with high precision.
**Accession Number:** WOS:000247008900003
**Author Identifiers:**

| Author | ResearcherID Number | ORCID Number |
|---|---|---|
| Orasan, Constantin | C-2677-2012 | 0000-0003-2067-8890 |

**ISSN:** 1076-9757
**eISSN:** 1943-5037

---

**Record 9 of 15**
**Title:** WordNet nouns: Classes and instances
**Author(s):** Miller, GA (Miller, GA); Hristea, F (Hristea, F)
**Source:** COMPUTATIONAL LINGUISTICS **Volume:** 32 **Issue:** 1 **Pages:** 1-3 **Published:** MAR 2006
**Abstract:** WordNet, a lexical database for English that is extensively used by computational linguists, has not previously distinguished hyponyms that are classes from hyponyms that are instances. This note describes an attempt to draw that distinction and proposes a simple way to incorporate the results into future versions of WordNet.
**Accession Number:** WOS:000237977900001
**ISSN:** 0891-2017

---

**Record 10 of 15**
**Title:** Evaluating WordNet-based measures of lexical semantic relatedness
**Author(s):** Budanitsky, A (Budanitsky, A); Hirst, G (Hirst, G)
**Source:** COMPUTATIONAL LINGUISTICS **Volume:** 32 **Issue:** 1 **Pages:** 13-47 **DOI:** 10.1162/coli.2006.32.1.13 **Published:** MAR 2006
**Abstract:** The quantification of lexical semantic relatedness has many applications in NLP, and many different measures have been proposed. We evaluate five of these measures, all of which use WordNet as their central resource, by comparing their performance in detecting and correcting real-word spelling errors. An information-content-based measure proposed by Jiang and Conrath is found superior to those proposed by Hirst and St-Onge, Leacock and Chodorow, Lin, and Resnik. In addition, we explain why distributional similarity is not an adequate proxy for lexical semantic relatedness.
**Accession Number:** WOS:000237977900003
**Author Identifiers:**

| Author | ResearcherID Number | ORCID Number |
|---|---|---|
| Hirst, Graeme | A-1825-2008 | |

**ISSN:** 0891-2017

---

**Record 11 of 15**
**Title:** Generative prior knowledge for discriminative classification
**Author(s):** Epshteyn, A (Epshteyn, Arkady); DeJong, G (DeJong, Gerald)
**Source:** JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH **Volume:** 27 **Pages:** 25-53 **Published:** 2006
**Abstract:** We present a novel framework for integrating prior knowledge into discriminative classifiers. Our framework allows discriminative classifiers such as Support Vector Machines (SVMs) to utilize prior knowledge specified in the generative setting. The dual objective of fitting the data and respecting prior knowledge is formulated as a bilevel program, which is solved ( approximately) via iterative application of second-order cone programming. To test our approach, we consider the problem of using WordNet ( a semantic database of English language) to improve low-sample classification accuracy of newsgroup categorization. WordNet is viewed as an approximate, but readily available source of background knowledge, and our framework is capable of utilizing it in a flexible way.
**Accession Number:** WOS:000240777600002
**ISSN:** 1076-9757

---

**Record 12 of 15**
**Title:** Comparing knowledge sources for nominal anaphora resolution
**Author(s):** Markert, K (Markert, K); Nissim, M (Nissim, M)
**Source:** COMPUTATIONAL LINGUISTICS **Volume:** 31 **Issue:** 3 **Pages:** 367-401 **DOI:** 10.1162/089120105774321064 **Published:** SEP 2005
**Abstract:** We compare two ways of obtaining lexical knowledge for antecedent selection in other-anaphora and definite noun phrase coreference. Specifically, we compare an algorithm that relies on links encoded in the manually created lexical hierarchy WordNet and an algorithm that mines corpora by means of shallow lexico-semantic patterns. As corpora we use the British National Corpus (BNC), as well as the Web, which has not been previously used for this task. Our results show that (a) the knowledge encoded in WordNet is often insufficient, especially for anaphor-antecedent relations that exploit subjective or context-dependent knowledge; (b) for other-anaphora, the Web-based method outperforms the WordNet-based method; (c) for definite NP coreference, the Web-based method yields results comparable to those obtained using WordNet over the whole data set and outperforms the WordNet-based method on subsets of the data set; (d) in both case studies, the BNC-based method is worse than the other methods because of data sparseness. Thus, in our studies, the Web-based method alleviated the lexical knowledge gap often encountered in anaphora resolution and handled examples with context-dependent relations between anaphor and antecedent. Because it is inexpensive and needs no hand-modeling of lexical knowledge, it is a promising knowledge source to integrate into anaphora resolution systems.
**Accession Number:** WOS:000232471600004
**ISSN:** 0891-2017

---

**Record 13 of 15**
**Title:** Learning domain ontologies from document warehouses and dedicated web sites
**Author(s):** Navigli, R (Navigli, R); Velardi, P (Velardi, P)
**Source:** COMPUTATIONAL LINGUISTICS **Volume:** 30 **Issue:** 2 **Pages:** 151-179 **DOI:** 10.1162/089120104323093276 **Published:** JUN 2004
**Abstract:** We present a method and a tool, OntoLearn, aimed at the extraction of domain ontologies from Web sites, and more generally from documents shared among the members of virtual organizations. OntoLearn first extracts a domain terminology from available documents. Then, complex domain terms are semantically interpreted and arranged in a hierarchical fashion. Finally, a general-purpose ontology, WordNet, is trimmed and enriched with the detected domain concepts. The major novel aspect of this approach is semantic interpretation, that is, the association of a complex concept with a complex term. This involves finding the appropriate WordNet concept for each word of a terminological string and the appropriate conceptual relations that hold among the concept components. Semantic interpretation is based on a new word sense disambiguation algorithm, called structural semantic interconnections.
**Accession Number:** WOS:000222533400002
**ISSN:** 0891-2017

---

**Record 14 of 15**
**Title:** CorMet: A computational, corpus-based conventional metaphor extraction system
**Author(s):** Mason, ZJ (Mason, ZJ)

**Source:** COMPUTATIONAL LINGUISTICS **Volume:** 30 **Issue:** 1 **Pages:** 23-44 **DOI:** 10.1162/089120104773633376 **Published:** MAR 2004
**Abstract:** CorMet is a corpus-based system for discovering metaphorical mappings between concepts. It does this by finding systematic variations in domain-specific selectional preferences, which are inferred from large, dynamically mined Internet corpora.
Metaphors transfer structure from a source domain to a target domain, making some concepts in the target domain metaphorically equivalent to concepts in the source domain. The verbs that select for a concept in the source domain tend to select for its metaphorical equivalent in the target domain. This regularity, detectable with a shallow linguistic analysis, is used to find the metaphorical interconcept mappings, which can then be used to infer the existence of higher-level conventional metaphors.
Most other computational metaphor systems use small, hand-coded semantic knowledge bases and work on a few examples. Although CorMet's only knowledge base is WordNet (Fellbaum 1998) it can find the mappings constituting many conventional metaphors and in some cases recognize sentences instantiating those mappings. CorMet is tested on its ability to find a subset of the Master Metaphor List (Lakoff, Espenson, and Schwartz 1991).
**Accession Number:** WOS:000220741000002
**ISSN:** 0891-2017

**Record 15 of 15**

**Title:** Automatic association of Web directories with word senses
**Author(s):** Santamaria, C (Santamaria, C); Gonzalo, J (Gonzalo, J); Verdejo, F (Verdejo, F)
**Source:** COMPUTATIONAL LINGUISTICS **Volume:** 29 **Issue:** 3 **Pages:** 485-502 **DOI:** 10.1162/089120103322711613 **Published:** SEP 2003
**Abstract:** We describe an algorithm that combines lexical information (from WordNet 1.7) with Web directories (from the Open Directory Project) to associate word senses with such directories. Such associations can be used as rich characterizations to acquire sense-tagged corpora automatically, cluster topically related senses, and detect sense specializations. The algorithm is evaluated for the 29 nouns (147 senses) used in the Senseval 2 competition, obtaining 148 (word sense, Web directory) associations covering 88% of the domain-specific word senses in the test data with 86% accuracy. The richness of Web directories as sense characterizations is evaluated in a supervised word sense disambiguation task using the Senseval 2 test suite. The results indicate that, when the directory/word sense association is correct, the samples automatically acquired from the Web directories are nearly as valid for training as the original Senseval 2 training instances. The results support our hypothesis that Web directories are a rich source of lexical information: cleaner, more reliable, and more structured than the full Web as a corpus.
**Accession Number:** WOS:000186631500006
**ISSN:** 0891-2017
**eISSN:** 1530-9312