CSE6390 3.0 Special Topics in AI & Interactive Systems II
Introduction to Computational Linguistics
Insructor: Nick Cercone – 3050 CSEB – nick@cse.yorku.ca
Tuesdays,Thursdays 10:00-11:30 – South Ross 104
Fall Semester, 2010

Word sense disambiguation

Edited from Wikipedia, the free encyclopedia

In computational linguistics (CL), word sense disambiguation (WSD) is an open problem of natural language processing, which comprises the process of identifying which sense of a word (i.e., meaning) is used in any given sentence, when the word has a number of distinct senses (polysemy). Solution of this problem impacts such other tasks of computation linguistics, such as discourse, improving relevance of search engines, anaphora resolution, coherence (linguistics), inference and others.

Research has progressed steadily to the point where WSD systems achieve consistent levels of accuracy on a variety of word types and ambiguities. A rich variety of techniques have been researched, from dictionary-based methods that use the knowledge encoded in lexical resources, to supervised machine learning methods in which a classifier is trained for each distinct word on a corpus of manually sense-annotated examples, to completely unsupervised methods that cluster occurrences of words, thereby inducing word senses. Among these, supervised learning approaches have been the most successful algorithms to date.

Current accuracy is difficult to state without a host of caveats. On English, accuracy at the coarse-grained (homograph) level is routinely above 90%, with some methods on particular homographs achieving over 96%. On finer-grained sense distinctions, top accuracies from 59.1% to 69.0% have been reported in recent evaluation exercises (SemEval-2007, Senseval-2), where the baseline accuracy of the simplest possible algorithm of always choosing the most frequent sense was 51.4% and 57%, respectively.

About

A disambiguation process requires two strict things: a dictionary to specify the senses which are to be disambiguated and a corpus of language data to be disambiguated. Also, WSD task has two variants: "all words" and "lexical sample" task. In the former, the program has to disambiguate all words, while the latter comprises of disambiguating only the words which were previously selected.

To give a hint how all this works, consider two examples of the distinct senses that exist for the (written) word "*bass*": (1) a type of fish and (2) tones of low frequency, and the sentences:

1. *I went fishing for some sea bass.*
2. *The bass line of the song is too weak.*

To a human, it is obvious that the first sentence is using the word "*bass*", as in the former sense above and in the second sentence, the word "*bass*" is being used as in the latter sense below. Developing algorithms to replicate this human ability can often be a difficult task.

History

WSD was first formulated as a distinct computational task for machine translation in the 1940s, making it one of the oldest CL problems. Warren Weaver, in his famous 1949 memorandum on translation[1], first introduced the problem in a computational context. Early researchers understood well the significance and difficulty of WSD. In fact, Bar-Hillel (1960) used the above example to argue that WSD could not be solved by "electronic computer" because of the need in general to model all world knowledge.

In the 1970s, WSD was a subtask of semantic interpretation systems developed within the field of artificial intelligence, but since WSD systems were largely rule-based and hand-coded they were prone to a knowledge acquisition bottleneck.

By the 1980s large-scale lexical resources, such as the Oxford Advanced Learner's Dictionary of Current English (OALD), became available: hand-coding was replaced with knowledge automatically extracted from these resources, but disambiguation was still knowledge-based or dictionary-based.

In the 1990s, the statistical revolution swept through computational linguistics, and WSD became a paradigm problem on which to apply supervised machine learning techniques.

The 2000s saw supervised techniques reach a plateau in accuracy, and so attention has shifted to coarser-grained senses, domain adaptation, semi-supervised and unsupervised corpus-based systems, combinations of different methods, and the return of knowledge-based systems via graph-based methods. Still, supervised systems continue to perform best.

Difficulties

Differency of dictionaries

One problem with word sense disambiguation is deciding what the senses are. In cases like the word *bass* above, at least some senses are obviously different. In other cases, however, the different senses can be closely related (one meaning being a metaphorical or metonymic extension of another), and in such cases division of words into senses becomes much more difficult. Different dictionaries and thesauruses will provide different divisions of words into senses. One solution some researchers have used is to choose a particular dictionary, and just use its set of senses. Generally, however, research results using broad distinctions in senses have been much better than those using narrow[2][3]. However, given the lack of a full-fledged coarse-grained sense inventory, most researchers continue to work on fine-grained WSD.

Most research in the field of WSD is performed by using WordNet as a reference sense inventory for English. WordNet is a computational lexicon that encodes concepts as synonym sets (e.g. the concept of car is encoded as { car, auto, automobile, machine, motorcar }). Other resources used for disambiguation purposes include Roget's Thesaurus and Wikipedia[4].

Part-of-speech tagging

In any real test POS-tagging and sense tagging are very closely related (it concerns only some languages, e.g. English) with each potentially making constraints to each other. And the question whether these tasks should be kept together or decoupled is still not unanimously resolved, but recently scientists incline to test these things separately (e.g., in the Senseval/Semeval competitions parts of speech are provided as input for the text to disambiguate).

It is instructive to compare the word sense disambiguation problem with the problem of part-of-speech tagging. Both involve disambiguating or tagging with words, be it with senses or parts of speech. However, algorithms used for one do not tend to work well for the other, mainly because the part of speech of a word is primarily determined by the immediately adjacent one to three words, whereas the sense of a word may be determined by words further away. The success rate for part-of-speech tagging algorithms is at present much higher than that for WSD, state-of-the art being around 95% accuracy or better, as compared to less than 75% accuracy in word sense disambiguation with supervised learning. These figures are typical for English, and may be very different from those for other languages.

Inter-judge variance

Another problem is inter-judge variance. WSD systems are normally tested by having their results on a task compared against those of a human. However, while it is relatively easy to assign parts of speech to text, training people to tag senses is far more difficult[5]. While users can memorize all of the possible parts of speech a word can take, it is impossible for individuals to memorize all of the senses a word can

take. Moreover, humans do not agree on the task at hand — give a list of senses and sentences, and humans will not always agree on which word belongs in which sense[6].

Thus, computer cannot be expected to give better performance on such a task than a human (indeed, since the human serves as the standard, the computer being better than the human is incoherent), so the human performance serves as an upper bound. Human performance, however, is much better on coarse-grained than fine-grained distinctions, so this again is why research on coarse-grained distinctions has been put to test in recent WSD evaluation exercises[2][3].

Common sense

Some AI researchers like Douglas Lenat argue that one cannot parse meanings from words without some form of common sense ontology. For example, comparing two these sentences:

- "Jill and Mary are sisters." — (they are sisters of each other).
- "Jill and Mary are mothers." — (each is independently a mother).

To properly identify senses of words one must know common sense facts[7]. Moreover, sometimes the common sense is needed to disambiguate such words like pronouns in case of having anaphoras or cataphoras in the text.

Sense inventory and algorithms' task-dependency

A task-independent sense inventory is not a coherent concept: each task requires its own division of word meaning into senses relevant to the task. For example, the ambiguity of 'mouse' (animal or device) is not relevant in English-French machine translation, but is relevant in information retrieval. The opposite is true of 'river', which requires a choice in French (fleuve 'flows into the sea', or rivière 'flows into a river').

Also, completely different algorithms might be required by different applications. In machine translation, the problem takes the form of target word selection. Here the "senses" are words in the target language, which often correspond to significant meaning distinctions in the source language (bank could translate to French banque 'financial bank' or rive 'edge of river'). In information retrieval, a sense inventory is not necessarily required, because it is enough to know that a word is used in the same sense in the query and a retrieved document; what sense that is, is unimportant.

Discreteness of senses

Finally, the very notion of "word sense" is slippery and controversial. Most people can agree in distinctions at the coarse-grained homograph level (e.g., pen as writing instrument or enclosure), but go down one level to fine-grained polysemy, and disagreements arise. For example, in Senseval-2, which used fine-grained sense distinctions, human annotators agreed in only 85% of word occurrences[8]. Word meaning is in principle infinitely variable and context sensitive. It does not divide up easily into distinct or discrete sub-meanings[9]. Lexicographers frequently discover in corpora loose and overlapping word meanings, and standard or conventional meanings extended, modulated, and exploited in a bewildering variety of ways. The art of lexicography is to generalize from the corpus to definitions that evoke and explain the full range of meaning of a word, making it seem like words are well-behaved semantically. However, it is not at all clear if these same meaning distinctions are applicable in computational applications, as the decisions of lexicographers are usually driven by other considerations. Recently, a task - named lexical substitution - has been proposed as a possible solution to the sense discreteness problem[10]. The task consists of providing a substitute for a word in context that preserves the meaning of the original word (potentially, substitutes can be chosen from the full lexicon of the target language, thus overcoming discreteness).

Approaches and methods

As in all natural language processing, there are two main approaches to WSD — deep approaches and shallow approaches.

Deep approaches presume access to a comprehensive body of; world knowledge. Knowledge, such as "you can go fishing for a type of fish, but not for low frequency sounds" and "songs have low frequency sounds as parts, but not types of fish", is then used to determine in which sense the word is used. These approaches are not very successful in practice, mainly because such a body of knowledge does not exist in a computer-readable format, outside of very limited domains. However, if such knowledge did exist, then deep approaches would be much more accurate than the shallow approaches. Also, there is a long tradition in computational linguistics, of trying such approaches in terms of coded knowledge and in some cases, it is hard to say clearly whether the knowledge involved is linguistic or world knowledge. The first attempt was that by Margaret Masterman and her colleagues, at the Cambridge Language Research Unit in England, in the 1950s. This attempt used as data a punched-card version of Roget's Thesaurus and its numbered "heads", as an indicator of topics and looked for repetitions in text, using a set intersection algorithm. It was not very successful[11], but had strong relationships to later work, especially Yarowsky's machine learning optimisation of a thesaurus method in the 1990s.

Shallow approaches don't try to understand the text. They just consider the surrounding words, using information such as "if *bass* has words *sea* or *fishing* nearby, it probably is in the fish sense; if *bass* has the words *music* or *song* nearby, it is probably in the music sense." These rules can be automatically derived by the computer, using a training corpus of words tagged with their word senses. This approach, while theoretically not as powerful as deep approaches, gives superior results in practice, due to the computer's limited world knowledge. However, it can be confused by sentences like *The dogs bark at the tree* which contains the word *bark* near both *tree* and *dogs*.

There are four conventional approaches to WSD:

- Dictionary- and knowledge-based methods: These rely primarily on dictionaries, thesauri, and lexical knowledge bases, without using any corpus evidence.
- Supervised methods: These make use of sense-annotated corpora to train from.
- Semi-supervised or minimally-supervised methods: These make use of a secondary source of knowledge such as a small annotated corpus as seed data in a bootstrapping process, or a word-aligned bilingual corpus.
- Unsupervised methods: These eschew (almost) completely external information and work directly from raw unannotated corpora. These methods are also known under the name of word sense discrimination.

Dictionary- and knowledge-based methods

The Lesk algorithm[12] is the seminal dictionary-based method. It is based on the hypothesis that words used together in text are related to each other and that the relation can be observed in the definitions of the words and their senses. Two (or more) words are disambiguated by finding the pair of dictionary senses with the greatest word overlap in their dictionary definitions. For example, when disambiguating the words in "pine cone", the definitions of the appropriate senses both include the words evergreen and tree (at least in one dictionary).

An alternative to the use of the definitions is to consider general word-sense relatedness and to compute the semantic similarity of each pair of word senses based on a given lexical knowledge base such as WordNet. Graph-based methods reminiscent of spreading activation research of the early days of AI research have been applied with some success. More complex graph-based approaches have been shown to perform almost as good as supervised methods[13].

The use of selectional preferences (or selectional restrictions) are also useful. For example, knowing that one typically cooks food, one can disambiguate the word bass in "I am cooking bass" (i.e., it's not a musical instrument).

Supervised methods

Supervised methods are based on the assumption that the context can provide enough evidence on its own to disambiguate words (hence, world knowledge and reasoning are deemed unnecessary). Probably every machine learning algorithm going has been applied to WSD, including associated techniques such

as feature selection, parameter optimization, and ensemble learning. Support vector machines and memory-based learning have been shown to be the most successful approaches, to date, probably because they can cope with the high-dimensionality of the feature space. However, these supervised methods are subject to a new knowledge acquisition bottleneck since they rely on substantial amounts of manually sense-tagged corpora for training, which are laborious and expensive to create.

Semi-supervised methods

The bootstrapping approach starts from a small amount of seed data for each word: either manually-tagged training examples or a small number of surefire decision rules (e.g., 'play' in the context of 'bass' almost always indicates the musical instrument). The seeds are used to train an initial classifier, using any supervised method. This classifier is then used on the untagged portion of the corpus to extract a larger training set, in which only the most confident classifications are included. The process repeats, each new classifier being trained on a successively larger training corpus, until the whole corpus is consumed, or until a given maximum number of iterations is reached.

Other semi-supervised techniques use large quantities of untagged corpora to provide co-occurrence information that supplements the tagged corpora. These techniques have the potential to help in the adaptation of supervised models to different domains.

Also, an ambiguous word in one language is often translated into different words in a second language depending on the sense of the word. Word-aligned bilingual corpora have been used to infer cross-lingual sense distinctions, a kind of semi-supervised system.

Unsupervised methods

Unsupervised learning is the greatest challenge for WSD researchers. The underlying assumption is that similar senses occur in similar contexts, and thus senses can be induced from text by clustering word occurrences using some measure of similarity of context[14]. Then, new occurrences of the word can be classified into the closest induced clusters/senses. Performance has been lower than other methods, above, but comparisons are difficult since senses induced must be mapped to a known dictionary of word senses. Alternatively, if a mapping to a set of dictionary senses is not desired, cluster-based evaluations (including measures of entropy and purity) can be performed. It is hoped that unsupervised learning will overcome the knowledge acquisition bottleneck because they are not dependent on manual effort.

Summary

Almost all these approaches normally work by defining a window of **N** content words around each word to be disambiguated in the corpus, and statistically analyzing those **N** surrounding words. Two shallow approaches used to train and then disambiguate are Naïve Bayes classifiers and decision trees. In recent research, kernel-based methods such as support vector machines have shown superior performance in supervised learning. Graph-based approaches, that currently achieve performance close to the state of the art, have also gained much attention from the research community.

Because of the lack of training data, many word sense disambiguation algorithms use semi-supervised learning, which allows both labeled and unlabeled data. The Yarowsky algorithm was an early example of such an algorithm[15]. It uses the ‘One sense per collocation’ and the ‘One sense per discourse’ properties of human languages for word sense disambiguation. From observation, words tend to exhibit only one sense in most given discourse and in a given collocation.

Evaluation of methods

The evaluation of WSD systems requires a test corpus hand-annotated with the target or correct senses, and assumes that such a corpus can be constructed. Two main performance measures are used:

- Precision: the fraction of system assignments made that are correct
- Recall: the fraction of total word instances correctly assigned by a system

If a system makes an assignment for every word, then precision and recall are the same, and can be called accuracy. This model has been extended to take into account systems that return a set of senses with weights for each occurrence.

There are two kinds of test corpora:

- Lexical sample: the occurrences of a small sample of target words need to be disambiguated, and
- All-words: all the words in a piece of running text need to be disambiguated.

The latter is deemed a more realistic form of evaluation, but the corpus is more expensive to produce because human annotators have to read the definitions for each word in the sequence every time they need to make a tagging judgement, rather than once for a block of instances for the same target word. In order to define common evaluation datasets and procedures, public evaluation campaigns have been organized. Senseval has been run three times: Senseval-1 (1998), Senseval-2 (2001), Senseval-3 (2004), and its successor, SemEval (2007), once.

References

1. W. Weaver. 1949. Translation. In Machine Translation of Languages: Fourteen Essays, ed. by Locke, W.N. and Booth, A.D. Cambridge, MA: MIT Press.
2. R. Navigli, K. Litkowski, O. Hargraves. 2007. SemEval-2007 Task 07: Coarse-Grained English All-Words Task. Proc. of Semeval-2007 Workshop (SEMEVAL), in the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, Czech Republic, pp. 30-35.
3. S. Pradhan, E. Loper, D. Dligach, M. Palmer. 2007. SemEval-2007 Task 17: English lexical sample, SRL and all words. Proc. of Semeval-2007 Workshop (SEMEVAL), in the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, Czech Republic, pp. 87-92.
4. R. Mihalcea. 2007. Using Wikipedia for Automatic Word Sense Disambiguation. In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL 2007), Rochester, April 2007.
5. C. Fellbaum 1997. Analysis of a handtagging task. Proceedings of ANLP-97 Workshop on Tagging Text with Lexical Semantics: Why, What, and How? Washington D.C., USA.
6. B. Snyder and M. Palmer. 2004. The English all-words task. In Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3), Barcelona, Spain, pp. 41-43.
7. D. Lenat. "Computers versus Common Sense". http://www.youtube.com/watch?v=KSrUHGaUE_c. Retrieved 2008-12-10. (GoogleTachTalks on youtube)
8. P. Edmonds. 2000. Designing a task for SENSEVAL-2. Tech. note. University of Brighton, Brighton. U.K.
9. A. Kilgarriff. 1997. I don't believe in word senses. Comput. Human. 31(2), pp. 91-113.
10. D. McCarthy, R. Navigli. 2009. The English Lexical Substitution Task, Language Resources and Evaluation, 43(2), Springer, pp. 139-159.
11. Y. Wilks, B. Slator, L. Guthrie. 1996. Electric Words: dictionaries, computers and meanings. Cambridge, MA: MIT Press.
12. M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. Proceedings of SIGDOC-86: 5th International Conference on Systems Documentation, Toronto, Canada, 24-26.
13. ^ R. Navigli, P. Velardi. 2005. Structural Semantic Interconnections: a Knowledge-Based Approach to Word Sense Disambiguation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 27(7), pp. 1063-1074.
14. H. Schütze. 1998. Automatic word sense discrimination. Computational Linguistics, 24(1), pp. 97-123.
15. D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pp. 189-196.

Suggested Reading

- Computational Linguistics Special Issue on Word Sense Disambiguation (1998)
- *Evaluation Exercises for Word Sense Disambiguation* The de-facto standard benchmarks for WSD systems.
- Roberto Navigli. *Word Sense Disambiguation: A Survey*, ACM Computing Surveys, 41(2), 2009, pp. 1-69. An up-to-date state of the art of the field.
- *Word Sense Disambiguation* as defined in Scholarpedia
- *Word Sense Disambiguation: The State of the Art* (PDF) A comprehensive overview By Prof. Nancy Ide & Jean Véronis (1998).
- *Word Sense Disambiguation Tutorial*, by Rada Mihalcea and Ted Pedersen (2005).

- *Word Sense Disambiguation: Algorithms and Applications*, edited by Eneko Agirre and Philip Edmonds (2006), Springer. Covers the entire field with chapters contributed by leading researchers. www.wsdbook.org site of the book
- Bar-Hillel, Yehoshua. 1964. Language and Information. New York: Addison-Wesley.
- Edmonds, Philip & Adam Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. Journal of Natural Language Engineering, 8(4):279-291.
- Edmonds, Philip. 2005. Lexical disambiguation. The Elsevier Encyclopedia of Language and Linguistics, 2nd Ed., ed. by Keith Brown, 607-23. Oxford: Elsevier.
- Ide, Nancy & Jean Véronis. 1998. Word sense disambiguation: The state of the art. Computational Linguistics, 24(1):1-40.
- Jurafsky, Daniel & James H. Martin. 2000. Speech and Language Processing. New Jersey, USA: Prentice Hall.
- Manning, Christopher D. & Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press. http://nlp.stanford.edu/fsnlp/
- Mihalcea, Rada. 2007. Word sense disambiguation. Encyclopedia of Machine Learning. Springer-Verlag.
- Resnik, Philip and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation, Natural Language Engineering, 5(2):113-133. http://www.cs.jhu.edu/~yarowsky/pubs/nle00.ps
- Yarowsky, David. 2000. Word sense disambiguation. Handbook of Natural Language Processing, ed. by Dale et al., 629-654. New York: Marcel Dekker.

Edited and retrieved from "http://en.wikipedia.org/wiki/Word_sense_disambiguation" (10 Jan 2010)