Word sense disambiguation

Philip Edmonds and Eneko Agirre (2008), Scholarpedia, 3(7):4358. doi:10.4249/scholarpedia.4358 revision #90370 [link to/cite this article]

- **Dr. Philip Edmonds**, Sharp Laboratories of Europe, Edmund Halley Road, Oxford Science Park, Oxford OX4 4GB, U.K.
- **Dr. Eneko Agirre**, University of the Basque Country, Donostia, Basque Country

In natural language processing, **word sense disambiguation** (WSD) is the problem of determining which "sense" (meaning) of a word is activated by the use of the word in a particular context, a process which appears to be largely unconscious in people. WSD is a natural classification problem: Given a word and its possible senses, as defined by a dictionary, classify an occurrence of the word in context into one or more of its sense classes. The features of the context (such as neighboring words) provide the evidence for classification.

A famous example is to determine the sense of *pen* in the following passage (Bar-Hillel 1960):

Little John was looking for his toy box. Finally he found it. The box was in the *pen*. John was very happy.

WordNet (http://wordnet.princeton.edu/) lists five senses for the word *pen*:

1. pen — a writing implement with a point from which ink flows.
2. pen — an enclosure for confining livestock.
3. playpen, pen — a portable enclosure in which babies may be left to play.
4. penitentiary, pen — a correctional institution for those convicted of major crimes.
5. pen — female swan.

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 (http://nlp.cs.swarthmore.edu/semeval/) , Senseval-2 (http://193.133.140.102/senseval2/)), where the baseline accuracy of the simplest possible algorithm of always choosing the most frequent sense was 51.4% and 57%, respectively.

Contents

History

WSD was first formulated as a distinct computational task during the early days of machine translation in the 1940s, making it one of the oldest problems in computational linguistics. Warren Weaver, in his famous 1949 memorandum on translation, 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.

Applications

Machine translation is the original and most obvious application for WSD but WSD has actually been considered in almost every application of language technology, including information retrieval, lexicography, knowledge mining/acquisition and semantic interpretation, and is becoming increasingly important in new research areas such as bioinformatics and the Semantic Web.

The utility of WSD

There is no doubt that the above applications require and use word sense disambiguation in one form or another. However, WSD as a separate module has not yet been shown to make a decisive difference in any application. There are a few recent results that show small positive effects in, for example, machine translation, but WSD has also been shown to hurt performance, as is the case in well-known experiments in information retrieval.

There are several possible reasons for this. First, the domain of an application often constrains the number of senses a word can have (e.g., one would not expect to see the 'river side' sense of *bank* in a financial application), and so lexicons can and have been constructed accordingly. Second, WSD might not be accurate enough yet to show an effect and moreover the sense inventory used is unlikely to match the specific sense distinctions required by the application. Third, treating WSD as a separate component or module may be misguided, as it might have to be more tightly integrated as an implicit process (i.e., as mutual disambiguation, below).

Machine translation

WSD is required for lexical choice in MT for words that have different translations for different senses. For example, in an English-French financial news translator, the English noun *change* could translate to either *changement* ('transformation') or *monnaie* ('pocket money'). However, most translation systems do not use a separate WSD module. The lexicon is often pre-disambiguated for a given domain, or hand-crafted rules are devised, or WSD is folded into a statistical translation model, where words are translated within phrases which thereby provide context.

Information retrieval

Ambiguity has to be resolved in some queries. For instance, given the query "depression" should the system return documents about illness, weather systems, or economics? Current IR systems (such as Web search engines), like MT, do not use a WSD module; they rely on the user typing enough context in the query to only retrieve documents relevant to the intended sense (e.g., "tropical depression"). In a process called mutual disambiguation, reminiscent of the Lesk method (below), all the ambiguous words are disambiguated by virtue of the intended senses co-occurring in the same document.

Information extraction and knowledge acquisition

In information extraction and text mining, WSD is required for the accurate analysis of text in many applications. For instance, an intelligence gathering system might need to flag up references to, say, illegal *drugs*, rather than medical *drugs*. Bioinformatics research requires the relationships between genes and gene products to be catalogued from the vast scientific literature; however, genes and their proteins often have the same name. More generally, the Semantic Web requires automatic annotation of documents according to a reference ontology. WSD is only beginning to be applied in these areas.

Methods

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 method (Lesk 1986) 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.

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. 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.

Evaluation

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 (http://www.itri.brighton.ac.uk/events/senseval/ARCHIVE/index.html) (1998), Senseval-2 (http://193.133.140.102/senseval2/) (2001), Senseval-3 (http://www.senseval.org/senseval3) (2004), and its successor, SemEval (http://nlp.cs.swarthmore.edu/semeval/) (2007), once.

Why is WSD hard?

This article discusses the common and traditional characterization of WSD as an explicit and separate process of disambiguation with respect to a fixed inventory of word senses. Words are typically assumed to have a finite and discrete set of senses, a gross simplification of the complexity of word meaning, as studied in lexical semantics. While this characterization has been fruitful for research into WSD per se, it is somewhat at odds with what seems to be needed in real applications, as discussed above.

WSD is hard for many reasons, three of which are discussed here.

A sense inventory cannot be task-independent

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').

Different algorithms for different applications

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.

Word meaning does not divide up into discrete 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. Word meaning is in principle infinitely variable and context sensitive. It does not divide up easily into distinct or discrete sub-meanings. 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.

References

Suggested reading

- Agirre, Eneko & Philip Edmonds (eds.). 2006. *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht: Springer. www.wsdbook.org (http://www.wsdbook.org/)
- 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.
- Lesk, Michael. 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.
- 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
- Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97-123.
- Weaver, Warren. 1949. Translation. In *Machine Translation of Languages: Fourteen Essays*, ed. by Locke, W.N. and Booth, A.D. Cambridge, MA: MIT Press.
- Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 189-196.http://www.cs.jhu.edu/~yarowsky/acl95.ps
- Yarowsky, David. 2000. Word sense disambiguation. *Handbook of Natural Language Processing*, ed. by Dale et al., 629-654. New York: Marcel Dekker.

Internal references

- Tomasz Downarowicz (2007) Entropy. Scholarpedia, 2(11):3901.
- Mark Aronoff (2007) Language. Scholarpedia, 2(5):3175.

External links

- Senseval website (http://www.senseval.org)
- SemEval website (http://nlp.cs.swarthmore.edu/semeval/)
- WSD tutorial (http://www.d.umn.edu/~tpederse/WSDTutorial.html)

See also

Linguistics, Natural Language Processing

Sponsored by: Prof. Ke CHEN, School of Computer Science, The University of Manchester, U.K.
Reviewed by (http://www.scholarpedia.org/w/index.php?title=Word_sense_disambiguation&oldid=32598) : Anonymous
Reviewed by (http://www.scholarpedia.org/w/index.php?title=Word_sense_disambiguation&oldid=40030) : Dr. Rada Mihalcea, University of North Texas
Accepted on: 2008-05-23 17:13:44 GMT (http://www.scholarpedia.org/w/index.php?title=Word_sense_disambiguation&oldid=40030)

Categories: Computational Intelligence | Pattern Recognition | Language, Speech, Music | Multiple Curators