

# Introduction to the Special Issue on Word Sense Disambiguation

Judita Preiss<sup>a,1</sup> and Mark Stevenson<sup>b</sup>

<sup>a</sup>*University of Cambridge, Computer Laboratory*  
Judita.Preiss@cl.cam.ac.uk

<sup>b</sup>*University of Sheffield, Department of Computer Science*  
M.Stevenson@dcs.shef.ac.uk

---

## 1 Introduction

Word sense disambiguation (WSD) is generally considered to be the task of automatically selecting relevant senses from a set of possibilities. Research into WSD has a long history in the field of Computational Linguistics and Natural Language Processing. During the pioneering days of early research into automatic translation in the 1950's it quickly became obvious that being able to identify the meanings of particular words would be necessary before further language processing could take place (Yngve, 1955). For example, an English-French machine translation system needs to know whether the word “bank” is being used to mean “financial institution” or “edge of river” to decide whether it should be translated as “banque” or “bord”. Many approaches to the WSD problem have been explored and, consequently, difficulties have also been brought to light. This special issue gathers together papers which discuss current obstacles and present the state of the art in WSD. This introduction is not intended to be a comprehensive survey of the state-of-the-art in WSD, instead it serves to place each of the works contained in this volume in some context. For surveys of WSD systems, see Ide and Véronis (1998) or Stevenson (2003), and, for more current systems, Preiss and Yarowsky (2002) or Mihalcea and Edmonds (2004).

It is possible to draw a distinction between final and intermediate tasks in language processing.<sup>2</sup> Final tasks carry out some task of interest for its own

---

<sup>1</sup> This work was supported by UK EPSRC project GR/N36462/93: ‘Robust Accurate Statistical Parsing (RASP)’.

<sup>2</sup> Spärck Jones and Galliers (1996) introduced a similar distinction, although they

sake; they are applications of potential interest to the non-linguist, and as such can be evaluated independently of any theory. Examples of final tasks include machine translation and summarisation – it is possible to judge automatically translated texts and automatically generated summaries without reference to linguistic concepts.

On the other hand, intermediate tasks may be employed by a system which carries out a final task, but the results of an intermediate task are not generally useful in themselves. WSD is an example of an intermediate task. Another example is parsing – the trees produced by a parser cannot be fully evaluated independently of a theoretical framework.

## 2 Linguistic assumptions of WSD

A number of linguistic assumptions are made to allow evaluation of WSD. Firstly, there is a choice over which lexicon is used to provide the set of potential meanings for each word. The different possible structures of the lexicon can have an effect on the form of the senses which are assigned to words. The traditional dictionary structure lists a set of possible senses for each word. Machine readable dictionaries (MRDs) such as LDOCE (Procter, 1978) use this structure. Some dictionaries provide a set of general categories, such as HUMAN or ANIMATE, which can be assigned to any word. One of the most commonly used lexicons for WSD research, WordNet, (five of the six papers in this special issue describe systems which use it) is like a dictionary in some aspects, since each word contains a list of possible senses (synsets). However, these synsets often apply to more than one lexical item, and in this respect WordNet is different to a traditional dictionary.

Using a dictionary which enumerates word senses is not without problems. For example, such a representation does not make clear the distinction between basic homonymy and logically related senses (Pustejovsky, 1995). Pustejovsky’s solution is to extend a core set of word senses and generate a lexicon. Most WSD approaches do require a predefined sense inventory, however it is not necessary in WSD based on sense clustering (e.g., Schütze (1998)), where instances are grouped together according to their occurring in the same sense.

Restricting ourselves to a predefined sense inventory, another linguistic assumption which must be made in WSD is whether a word should be assigned exactly one sense or whether assigning multiple senses is appropriate. It has generally been assumed that each word should be assigned exactly one sense,

---

use the terms “systems” and “components” which are similar to final and intermediate tasks respectively.

although there have been arguments for assigning multiple senses. Even for humans the task of selecting appropriate senses from a given list is non-trivial (Véronis, 1998): it is in no way clear that people obtain the meaning of words by comparing them to a list of senses, and ambiguity tests are not always infallible (Zwicky and Sadock, 1975). Indeed, even between expert lexicographers, the inter-annotator agreement can vary hugely (Ahlsweide and Lorand, 1993), (Kilgarriff and Rosenzweig, 2000b), although the granularity of the chosen sense inventory (such as a dictionary) plays an important role in the level of agreement attained (Hanks, 2003), (Atkins and Levin, 1991).

Murray and Green’s paper, “Lexical knowledge and human disagreement on a WSD task”, addresses the lack of inter-annotator agreement, reaching the conclusion that “*agreement serves to describe the relative knowledge of the judges far more than it describes the “correctness” of their judgements*”.

### 3 Evaluation of WSD

A major recent advance in the field has been the creation of the SENSEVAL evaluation exercises, which provide a uniform framework for comparing the performance of WSD systems. As with any standardised evaluation framework, this approach to evaluation has some disadvantages. Firstly, choices are made for the linguistic assumptions within the framework which may not suit all approaches. A second disadvantage is that there is a danger that the field is led into an evaluation-led research agenda in which the goal of reporting ever improved results on a particular test set is given undue prominence. However, the overall effect of the SENSEVAL exercises has been highly positive for WSD research. The meetings have served as an international focus for WSD research and the evaluation materials created for the exercises provide a valuable resource for researchers where previously there had been a severe shortage of suitable test material. Systems submitted to SENSEVAL included both unsupervised (systems not requiring training data) and supervised systems, and the result is a precision/recall ranking of participating systems for each task (see the SENSEVAL-1 results (Kilgarriff and Rosenzweig, 2000a), the SENSEVAL-2 proceedings (Preiss and Yarowsky, 2002), and the SENSEVAL-3 proceedings (Mihalcea and Edmonds, 2004) for more information on the individual tasks and participating systems).

Using a predefined sense inventory and comparing answers against a gold standard is still the most frequent method for evaluating WSD, however alternatives have been explored. For example, the later SENSEVAL exercises have introduced some application-based evaluations of WSD (i.e., when WSD is used within a final task). In this case, the sense inventory in any application-based evaluation method is dictated by the application, avoiding the potentially very

fine-grained sense splits present in some current dictionaries (Kilgariff, 1997).

As an intermediate task the ultimate usefulness of WSD will be determined by whether it can be used to improve the performance of a final system. In this Special Issue Véronis' paper on "HyperLex: Lexical Cartography for Information Retrieval" is an example of an application based evaluation of WSD. One of the important contributions of this work is to demonstrate that the HyperLex system can be used to enhance the performance of an information retrieval system.

## 4 Approaches to WSD

There are common approaches to some intermediate tasks in Natural Language Processing; the noisy channel model is often used for part of speech tagging (Church, 1988) and two-level morphology is accepted as standard (Koskeniemi, 1984). But, despite the fact that WSD has been an area of research for around half a century, there is still no agreement over the best approach to the problem. There have, however, been some general trends in WSD research and it is likely that there will eventually lead to a shared approach.

- For gold-standard based evaluations, large-scale lexicons are now commonly used to provide the set of possible senses associated with each word. This is in contrast to historical systems which often used small hand-crafted lexicons containing a wide variety of different linguistic knowledge sources (e.g., (Wilks, 1972) or (Hirst, 1987)). These extended lexicons have allowed a shift in the field from systems which disambiguate a small set of sample words to ones which can provide meanings for all content words in text. These large-scale systems will be more useful for final tasks such as machine translation, than the systems which can disambiguate only a few lexical items.
- The availability of large-scale lexicons has provided access to a wide variety of linguistic knowledge sources which can be used to inform the disambiguation process. For example, MRDs commonly include dictionary definitions, thesauruses contain groups of topically-related words and WordNet provides an ontologically-motivated hierarchy. WSD researchers have made use of the linguistic knowledge available to them which generally depends on the lexicon being used. The use of large-scale lexicons has provided access to a wide variety of knowledge sources which can be used in the disambiguation process. Disambiguation algorithms are now often driven by the type of linguistic information available in the lexical resource which provides the employed set of senses. For example, Agirre and Rigau (1996) made use of the hierarchical structure of WordNet, Yarowsky (1992) used the thesaural topics in Roget's Thesaurus, Bruce and Guthrie (1992) used LDOCE

subject codes while Lesk (1986) used dictionary definitions.

- In addition to large-scale lexicons, researchers now have access to increasing larger text corpora, including the internet, and this has proved useful information for WSD systems. However, there is a limited amount of available text which is annotated with senses (Ng, 1997). The corpora from the two SENSEVAL exercises are now commonly used but are still small compared to the unannotated ones which are now commonly used (Banko and Brill, 2001; Curran and Moens, 2002). An innovative approach to this lack of suitable corpora has been to make use of parallel text (Brown et al., 1991). Distinct senses of a word are often translated differently. For example, the translation of “bank” as “bord” would indicate that it was being used to mean “edge or river”. Since the translations can be identified in parallel text it is possible to automatically identify the sense in which the word is being used. However, like sense-annotated text, large parallel texts have also proved difficult to obtain.
- The combination of linguistic knowledge sources or disambiguation algorithms has emerged as a common technique for tackling the WSD problem. This approach dates back, at least, to Hirst’s “polaroid words” (Hirst, 1987) which relied on several types of information from a hand-crafted knowledge base. However, more recent approaches, such as McRoy (1992), Ng and Lee (1996) and Stevenson and Wilks (2001) have combined several of the knowledge sources available in large-scale lexicons.
- Another common theme which runs through recent WSD research has been the use of techniques from machine learning. These approaches have often been deployed to make the best use of the linguistic knowledge available to the WSD systems. Examples include Yarowsky (1993) who used decision lists to choose the collocation which is most likely to identify the correct sense for a word and Véronis and Ide (1990) performed WSD using a neural network automatically generated from a MRD. Others, including Ng and Lee (1996) and Stevenson and Wilks (2001), have used learning techniques to combine several knowledge sources. The performance of these combined systems was higher than the accuracy of any individual knowledge source used alone.

These general trends have provided a direction for WSD research and their influence is apparent in the remainder of the papers in this special issue. Seo et al.’s paper, “Unsupervised Word Sense Disambiguation using WordNet Relatives”, presents an unsupervised system based on the hierarchical structure of the WordNet dictionary (Miller et al., 1990), along the lines of Yarowsky (1992), and make use of the SENSEVAL corpora to put their results into context of other related works. Instead of using a sense annotated corpus, and thus creating a supervised system, the authors’ sense selection is based on conceptual distance (e.g., Agirre and Rigau (1996)) arising from frequency co-occurrence in untagged corpora.

In “Unsupervised and Supervised Exploitation of Semantic Domains in Lexical Disambiguation”, Gliozzo et al. formally model domains for WSD, and use the WordNet extension WordNet Domains (Magnini and Cavaglia, 2000) to create a number of WSD systems.

Further heuristical WSD approaches are presented in Moldovan and Novischi’s paper, “Word Sense Disambiguation of WordNet glosses”; in this work, WSD is employed to disambiguate WordNet glosses. Such an enhanced hierarchy has been shown to be more useful for future WSD applications (Harabagiu and Moldovan, 1998).

Preiss’ paper, “Probabilistic Word Sense Disambiguation”, is an example of the combination of knowledge sources using a combination of supervised and unsupervised systems. The work is motivated by the increase in performance when a number of WSD approaches are combined (Stevenson, 2003), and presents a novel theoretically motivated method for combining numerous approaches which are known to be successful.

## 5 Summary

This special issue presents work which questions the fundamental points of the field of WSD, such as annotation and evaluation (Murray and Green). A possible solution to evaluation of WSD systems is suggested in the form of an application based evaluation using HyperLex (Véronis). A number of state-of-the-art systems are presented; Moldovan and Novischi show how to extend further current knowledge, so it can be used by future WSD systems. Two systems making use of existing resources are presented (Seo et al., Gliozzo et al.), and a more theoretical approach to known methods is described by Preiss.

## References

- Agirre, E., Rigau, G., 1996. Word sense disambiguation using conceptual density. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING-96). Copenhagen, Denmark, pp. 16–22.
- Ahlsvede, T., Lorand, D., 1993. Word sense disambiguation by human subjects: Computational and psycholinguistic implications. In: Proceedings of the Workshop on Acquisitions of Lexical Knowledge from Text. Columbus, Ohio, pp. 1–9.
- Atkins, B. T. S., Levin, B., 1991. Admitting impediments. In: Zernik, U. (Ed.), Lexical acquisition. Lawrence Erlbaum, pp. 233–262.

- Banko, M., Brill, E., 2001. Scaling to very very large corpora for natural language disambiguation. In: Proceedings of 39th Annual Meeting of the Association for Computational Linguistics (ACL-01). Toulouse, France, pp. 26–33.
- Brown, P., Pietra, S. D., Pietra, V. D., Mercer, R., 1991. Word sense disambiguation using statistical methods. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91). Berkley, CA., pp. 264–304.
- Bruce, R., Guthrie, L., 1992. Genus disambiguation: A study in weighted preference. In: Proceedings of the 14th International Conference on Computational Linguistics (COLING-92). Nantes, France, pp. 1187–1191.
- Church, K., 1988. A stochastic parts program and noun phrase parser for unrestricted text. In: Proceedings of the Second Conference on Applied Natural Language Processing (ANLP-88). Austin, TX, pp. 136–143.
- Curran, J., Moens, M., 2002. Scaling context space. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02). University of Philadelphia, Pennsylvania, pp. 231–328.
- Hanks, P., 2003. Lexicography. In: Mitkov, R. (Ed.), *The Oxford handbook of computational linguistics*. Oxford University Press, pp. 48–69.
- Harabagiu, S., Moldovan, D., 1998. Knowledge processing on an extended WordNet. In: Fellbaum, C. (Ed.), *WordNet – An Electronic Lexical Database*. MIT Press, pp. 379–406.
- Hirst, G., 1987. *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge University Press.
- Ide, N., Véronis, J., 1998. The state of the art. *Computational Linguistics* 24, 1–40, Introduction to the Special Issue on Word Sense Disambiguation.
- Kilgarriff, A., 1997. “I don’t believe in word senses”. *Computers and the Humanities* 31 (2), 91–113.
- Kilgarriff, A., Rosenzweig, J., 2000a. English SENSEVAL: Report and results. In: Proceedings of Second International Conference on Language Resources and Evaluation (LREC-2000). Athens, Greece, pp. 1239–1244.
- Kilgarriff, A., Rosenzweig, J., 2000b. Framework and results for English SENSEVAL. *Computers and the Humanities* 34 (1–2), 15–48.
- Koskenniemi, K., 1984. A general computational model for word-form recognition and production. In: Proceedings of the Tenth International Conference on Computational Linguistics (COLING-84). Stanford, CA, pp. 178–181.
- Lesk, M., 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of AGM SIGDOC Conference. pp. 24–26.
- Magnini, B., Cavaglià, G., 2000. Integrating subject field codes into wordnet. In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000). Athens, Greece, pp. 1413–1418.
- McRoy, S., 1992. Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics* 18 (1), 1–30.
- Mihalcea, R., Edmonds, P. (Eds.), 2004. *Proceedings of SENSEVAL-3*. Forth-

- coming.
- Miller, G., Beckwith, R., Felbaum, C., Gross, D., Miller, K., 1990. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography* 3 (4), 235–244.
- Ng, H., 1997. Getting serious about word sense disambiguation. In: *Proceedings of the SIGLEX Workshop Tagging Text with Lexical Semantics*. Washington, DC, pp. 1–7.
- Ng, H., Lee, H., 1996. Integrating multiple knowledge sources to disambiguate word sense. an exemplar-based approach. In: *Proceedings of the 34th Meeting of the Association for Computational Linguistics (ACL-96)*. Santa Cruz, CA, pp. 40–47.
- Preiss, J., Yarowsky, D. (Eds.), 2002. *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguating Systems*.
- Procter, P., 1978. *Longman Dictionary of Contemporary English*. Longman Group Ltd.
- Pustejovsky, J., 1995. *The Generative Lexicon*. MIT Press.
- Schütze, H., 1998. Automatic word sense discrimination. *Computational Linguistics* 24 (1), 97–124.
- Spärck Jones, K., Galliers, J., 1996. *Evaluating Natural Language Processing Systems, an Analysis and Overview*. Springer Verlag, Lecture Notes in Artificial Intelligence 1083.
- Stevenson, M., 2003. *Word Sense Disambiguation: The Case for Combining Knowledge Sources*. CSLI Publications, Stanford, CA.
- Stevenson, M., Wilks, Y., 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics* 27 (3), 321–349.
- Véronis, J., 1998. A study of polysemy judgements and inter-annotator agreement. In: *Programme and advanced papers of the SENSEVAL workshop*. Herstmonceux Castle, UK, pp. 2–4.
- Véronis, J., Ide, N., 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In: *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*. Helsinki, Finland, pp. 389–394.
- Wilks, Y., 1972. *Grammar, Meaning and the Machine Analysis of Language*. Routledge, London.
- Yarowsky, D., 1992. Word sense disambiguation using statistical models of Roget’s categories trained on large corpora. In: *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*. Vol. 14. Nantes, France, pp. 454–460.
- Yarowsky, D., 1993. One sense per collocation. In: *Proceedings ARPA Human Language Technology Workshop*. Princeton, NJ, pp. 266–271.
- Yngve, V. H., 1955. Syntax and the problem of multiple meaning. In: Locke, W. N., Booth, A. D. (Eds.), *Machine translation of languages*. John Wiley and Sons, New York, pp. 208–226.
- Zwicky, A. M., Sadock, J. M., 1975. Ambiguity tests and how to fail them. *Syntax and Semantics* 4, 1–36.