Published in the ELRA Newsletter, Vol. 7 No. 3, 2002.

SENSEVAL: The evaluation of word sense disambiguation systems

Philip Edmonds, Sharp Laboratories of Europe, Oxford Science Park, Oxford OX4 4GB
phil@sharp.co.uk
2 Oct 2002

1 Word sense disambiguation

Word sense disambiguation (WSD) is the problem of deciding which sense a word has in any given context. The problem of doing WSD by computer is not new; it goes back to the early days of machine translation. But like other areas of computational linguistics, research into WSD has seen a resurgence because of the availability of large corpora. Statistical methods for WSD, especially techniques in machine learning, have proved to be very effective, as SENSEVAL has shown us.

In many ways, WSD is similar to part-of-speech tagging. It involves labelling every word in a text with a tag from a pre-specified set of tag possibilities for each word by using features of the context and other information. Like part-of-speech tagging, no one really cares about WSD as a task on its own, but rather as part of a complete application in, for instance, machine translation or information retrieval. Thus, WSD is often fully integrated into applications and cannot be separated out (for instance, in information retrieval WSD is often not done explicitly but is just by-product of query to document matching). But in order to study and evaluate WSD, researchers have concentrated on standalone, generic systems for WSD. This article is not about methods or uses of WSD, but about evaluation.

2 SENSEVAL

The success of any project in WSD is clearly tied to the evaluation of WSD systems. **SENSEVAL** was started in 1997, following a workshop, "Tagging with Lexical Semantics: Why, What, and How?", held at the conference on Applied Natural Language Processing. It's mission is to organise and run evaluation and related activities to test the strengths and weaknesses of WSD systems with respect to different words, different aspects of language, and different languages. It's underlying goal is to further our understanding of lexical semantics and polysemy.

SENSEVAL is run by small elected committee under the auspices of ACL-SIGLEX (the special interest group on the lexicon of the Association for Computational Linguistics). It is independent from other evaluation programmes in the language technology community, such as TREC and MUC, and, as yet, receives no permanent funding.

SENSEVAL held its first evaluation exercise in the summer of 1998, culminating in a workshop at Herstmonceux Castle, England on September 2–4 (Kilgarriff and Palmer 2000). Following the success of the first workshop, SENSEVAL-2, supported by EURALEX, ELSNET, EPSRC, and ELRA, was organized in 2000–2001. The Second International Workshop on Evaluating Word Sense Disambiguation Systems was held in conjunction with ACL-2001 on July 5–6, 2001 in Toulouse (Preiss and Yarowsky 2001).

The rest of this article describes the SENSEVAL-2 exercise—it's tasks, participants, scoring, and results. The article concludes with a short discussion of where SENSEVAL is heading.

3 SENSEVAL-2: Tasks and participants

The main goal of SENSEVAL-2 was to encourage new languages to participate, and to develop a methodology for all-words evaluation. We were successful: SENSEVAL-2 evaluated WSD systems on three types of task on 12 languages as follows:

All-words	Czech, Dutch, English, Estonian
Lexical sample	Basque, English, Italian, Japanese, Korean, Spanish, Swedish
Translation	Japanese

In the **all-words** task, systems must tag almost all of the content words in a sample of running text. In the **lexical sample** task, we first carefully select a sample of words from the lexicon; systems must then tag several instances of the sample words in short extracts of text. The **translation** task (Japanese only) is a lexical sample

task in which word sense is defined according to translation distinction. (By contrast, SENSEVAL-1 evaluated systems on only lexical sample tasks in English, French, and Italian.)

Table 1 gives a breakdown of the number of submissions and teams who participated in each task. Overall, 93 systems were submitted from 34 different teams. Some teams submitted multiple systems to the same task, and some submitted systems to multiple tasks. Dutch data was also prepared, but was not available in the exercise. Inter-annotator agreement (IAA), and system performance is discussed below.

Language	Task	No. of submissions	No. of teams	IAA	Baseline	Best system
Czech	AW	1	1	-	-	.94
Basque	LS	3	2	.75	.65	.76
Estonian	AW	2	2	.72	.85	.67
Italian	LS	2	2	-	-	.39
Korean	LS	2	2	-	.71	.74
Spanish	LS	12	5	.64	.48	.65
Swedish	LS	8	5	.95	-	.70
Japanese	LS	7	3	.86	.72	.78
Japanese	TL	9	8	.81	.37	.79
English	AW	21	12	.75	.57	.69
English	LS	26	15	.86	.51/.16	.64/.40

Table 1 Submissions to SENSEVAL-2

A task in SENSEVAL consists of three types of data: 1) A sense inventory of word-to-sense mappings, with possibly extra information to explain, define, or distinguish the senses (e.g., WordNet); 2) A corpus of manually tagged text or samples of text that acts as the Gold Standard, and that is split into an optional training corpus and test corpus; and 3) An optional sense hierarchy or sense grouping to allow for fine or coarse grained sense distinctions to be used in scoring. General guidelines for designing tasks were issued to ensure common evaluation standards (Edmonds 2000), but each task was designed individually.

WordNet was used for the first time in SENSEVAL; version 1.7 for the English tasks, and versions of EuroWordNet for Spanish, Italian, and Estonian. WordNet was chosen because of its wide availability and broad coverage, despite the often unmotivated demarcation of senses (Wordnet was designed from the point of view of synonymy rather than polysemy). In fact, WordNet 1.7 now includes revisions suggested by the human-tagging exercise for SENSEVAL-2.

The **Gold Standard corpus** must be replicable; the goal is to have human annotators agree at least 90% of the time. In practice, agreement was lower (see Table 1). At least two human annotators were required to tag every instance of a word, but often more annotators were involved in order to settle disagreements.

4 SENSEVAL-2: Evaluation procedure and results

Regardless of the type of task, each system is required to tag the words specified in the test corpus with one or more tags in the sense inventory, giving probabilities (or confidence values) if desired. A distinction is made between **supervised** systems, that use the training corpus, and **unsupervised** systems, that do not. An orthogonal distinction is made between systems that use just the test corpus (pure unsupervised) and systems that use other knowledge sources, such as dictionaries or corpora, but, in practice, few systems are pure.

The evaluation was run centrally from a single website at the University of Pennsylvania and followed the same procedure as used in SENSEVAL-1. For each task, data was released in three stages: trial data, training data (if available), and test data. Each team registered their system, and then downloaded the required data according to a set schedule. Teams had 21 days to work with the training data and 7 days with the test data. Each team submitted their answers to the website for automatic scoring. (The Japanese tasks were handled separately because of copyright issues.)

SENSEVAL-1 established a scoring system that was used again in SENSEVAL-2 with little change. Fine-grained scoring was used to score all systems. If the task had a sense hierarchy or grouping, then coarse-grained scoring was also done. In fine-grained scoring, a system had to give at least one of the Gold Standard senses. In coarse-grained scoring, all senses in the answer key and in system output are collapsed to their highest parent or

group identifier. For sense hierarchies, mixed-grained scoring was also done: a system is given partial credit for choosing a sense that is a parent of the required sense according to Melamed and Resnik's (1997) scheme.

Systems are not required to tag all instances of a word, or even all words, thus, precision and recall can be used, although the measures are not completely analogous to IR evaluation. **Recall** (percentage of right answers on all instances in the test set) is the basic measurement of accuracy in this task, because it shows how many correct disambiguations the system achieved overall. **Precision** (percentage of right answers in the set of answered instances) favours systems that are very accurate if only on a small subset of cases that the system could give answers to. **Coverage,** the percentage of instances that a system gives any answer to, is also reported.

Table 1 gives an overview of the results, as reported in Preiss and Yarowsky (2001). Inter-annotator agreement (generally, the percentage of cases where two human annotators agree on a sense, but this varies depending on the task), is shown. Baseline performance is generated in different ways, but usually as most frequent sense in the tagged corpus. The recall of the best system with perfect or near-perfect coverage is given for each task. For the English lexical sample task, scores for supervised and unsupervised systems are separated by a slash.

Notably, the results in SENSEVAL-2 were about 14 percentage points lower than in SENSEVAL-1 (for the English lexical sample), even though the same evaluation methodology was used and many of systems were improved versions of the same systems that participated in SENSEVAL-1. This can be seen as evidence that WordNet sense distinctions are indeed not well-motivated, but more research is required to confirm this.

Edmonds (2001) gives a more complete account of SENSEVAL-2 evaluation methodology. Almost all data and results of SENSEVAL is in the public domain. Visit the website to download it.

5 Where next?

SENSEVAL-2 was very successful in opening up new avenues for research into WSD and polysemy. It's clear that the current best systems achieve their high performance by using supervised machine learning. Research is now ongoing to explore how feature selection for the machine learning algorithms affects the performance on different types of polysemy. Indeed, it is hoped that we can now identify different types of polysemy on the basis of how easy or difficult the words are to disambiguate with different features and methods. Another result of SENSEVAL-2 was to underline the importance of a well-motivated sense inventory with the right level of granularity of sense distinction. If humans cannot reliably disambiguate a word based on the information in the sense inventory, then there is no meaningful way of evaluating a system. Efforts are ongoing to design new methodologies for building sense inventories and for annotating large corpora, which will inform research in lexicographics and lexical semantics. In particular, researchers are investigating methods to form well-motivated groupings of senses. Finally, the task of WSD set up in SENSEVAL is very divorced from real applications. Questions run from whether the sense distinctions in generic resources are useful in particular applications or domains, to whether a separate WSD module is useful, to whether we need to make explicit sense distinctions at all.

Planning for SENSEVAL-3 is currently underway and the SENSEVAL Committee welcomes proposals for tasks to be run as part of exercise. Any task that can test a word sense disambiguation (WSD) system, be it application dependent or independent, will be considered. The committee especially encourages tasks for different languages, cross-lingual tasks, and tasks that are relevant to particular NLP applications such as MT and IR. It also encourages tasks for areas related to WSD such as semantic tagging and domain classification.

Visit http://www.senseval.org/ for more details.

6 References

Philip Edmonds (2000). *Designing a task for SENSEVAL-2.* Technical Note. SENSEVAL-2 website.

Philip Edmonds and Scott Cotton (2001). SENSEVAL-2: Overview. In Preiss and Yarowksy (2001), pages 1–5.

Adam Kilgarriff and Martha Palmer (2000). Guest editors. Special Issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs. *Computers and the Humanities 34(1–2).*

Dan Melamed and Phil Resnik (2000). Tagger evaluation given hierarchical tag sets. *Computers and the Humanities 34(1–2).*

Judita Preiss and David Yarowsky (2001). Editors. *The Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems.*

SENSEVAL Website: http://www.senseval.org/