Sense Tag Annotation of Multilingual Corpora

Deliverable D5.2
ISLE Computational Lexicon Working Group

Authors

Philip Edmonds[1], Scott Cotton[2], Martha Palmer[2]

1: Sharp Laboratories of Europe, Oxford, UK
2: Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA

ISLE Deliverable D5.2: Sense Tag Annotation of Multilingual Corpora

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

Scott Cotton, Martha Palmer
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
cotton, mpalmer@linc.cis.upenn.edu

Abstract

ISLE Workpackage 5, for the Computational Lexicon Working Group, includes Task 5.2: the annotation of a sample corpus with sense tags. The execution of this task was carried out in concert with SENSEVAL-2: The Second International Workshop on Evaluating Word Sense Disambiguation Systems, which was held on July 5-6, 2001. This paper gives an overview of the creation of the sense tagged data, discussing the evaluation exercise, the tasks, the scoring system, the results, and information about the availability of the annotated corpora. A companion paper is appended, Making Fine-grained and Coarse-grained sense distinctions, both manually and automatically, which has been accepted by the *Journal of Natural Language Engineering* and is currently under revision. The authors are Martha Palmer, Hoa Dang and Christiane Fellbaum, and this paper addresses the issue of sense granularity, one of the topics for Deliverable 5.2.

1 Introduction

Highly ambiguous words pose continuing problems for Natural Language Processing (NLP) applications. They can lead to irrelevant document retrieval in IR systems, and inaccurate translations in Machine Translation systems. Several efforts have been made to develop automatic Word Sense Disambiguation (WSD) systems that are capable of addressing these problems (Ide and Veronis, 1998, Palmer and Light, 1999). While homonyms[1] like *bank* are fairly tractable, polysemous words like *run*, with related but subtly distinct meanings, present the greatest hurdle for WSD. The most polysemous words are not only the most frequently occurring ones, but many of their senses are also domain-independent, making the WSD problem ubiquitous. Attempts are currently being made to create corpora annotated with sense tags to enable the training of supervised WSD systems, as reported here.

SENSEVAL was started in 1997, under the auspices of ACL-SIGLEX, to bring together researchers to discuss and solve the WSD-evaluation problem. Its aim is to evaluate the strengths and weaknesses of WSD algorithms and systems with respect to different words, different varieties of language, and different languages.

SENSEVAL is independent from other evaluation programs in the language technology community, such as TREC and MUC. Unlike these programs, SENSEVAL is a 'freelance' program that is run entirely by volunteers. We'd like to remind everyone that while

[1] Word forms with multiple unrelated meanings.

SENSEVAL takes the guise of a competition, its main function is not to determine a winner but to explore the scientific aspects of word sense disambiguation.

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, ELRA, and ISLE 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. This paper gives an overview of SENSEVAL-2, discussing the evaluation exercise, the tasks, the scoring system, and the results. It ends with some recommendations for future evaluation exercises.

SENSEVAL2 provided a suitable framework for the execution of one of the ISLE Computational Lexicon Working Group tasks, Task 5.2: the annotation of a sample corpus with sense tags. The original Workpackage suggested the use of SIMPLE entries for this task, but during the progress of the project it was decided that EuroWordNet entries, which could be mapped to SIMPLE entries, were the most publicly available. This allowed the SENSEVAL2 sense tagged data to be made widely available and encouraged the broadest possible participation. The task of analyzing the tagged data in its multilingual context is still ongoing.

2 Tasks and participants

A main goal of SENSEVAL-2 was to encourage new languages to participate. 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. Task design is discussed in section 0 below.

93 systems were submitted from 34 different research teams. Table 1 gives a breakdown of the number of submissions and teams who participated in each task. Note that some teams submitted multiple systems to the same task, and some submitted systems to multiple tasks. Several tasks had no submissions: the Chinese and Danish tasks could not find enough time to complete the data in time for the exercise, and the available Dutch data was misplaced in the process of making it public. The Dutch data is available, and the Chinese and Danish data will be prepared in due course.

Language	Task	No. of	No.

		submissions	of teams
Chinese	LS	0	0
Danish	LS	0	0
Dutch	AW	0	0
Czech	AW	1	1
Basque	LS	3	2
Estonian	AW	2	2
Italian	LS	2	2
Korean	LS	2	2
Spanish	LS	12	5
Swedish	LS	8	5
Japanese	LS	7	3
Japanese	TL	9	8
English	AW	21	12
English	LS	26	15
Total		93	57

Table 1 Submissions to SENSEVAL-2

3 Task design

A task in SENSEVAL consists of three types of data: 1) A lexicon 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. 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 lexicon. Supervised systems can train on the training corpus, if one is available.

The SENSEVAL committee issued general guidelines for designing a task (Edmonds 2000). But it was up to the individual task organisers, to design their own tasks since each had different constraints on resource availability (both human and data). Everyone, however, used a common XML data encoding format developed for SENSEVAL-2. Specific issues in choosing and designing the resources for each task are described in the papers in this proceedings, and, more generally, by Kilgarriff and Rosenzweig (2000).

3.1 Lexicon and lexical samples

Each task organiser chose the lexicon for their task. Notably, 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. For the lexical sample tasks, the guidelines suggests that words be chosen from different parts of speech, different frequencies in the corpus, and different polysemies (i.e., number of senses). The number of words depended on the available resources. The sample words were kept secret from the wider community until the training data was released; however, the organisers consulted each other so that translations of some of the sample words could be used across tasks.

To give an example, the English Lexical Sample task used English WordNet as its sense inventory, and involved 73 lexical items taken from WordNet 1.7. WordNet 1.7 is a large electronic database organized as a semantic network built on paradigmatic relations like synonymy, hyponymy, antonymy, and entailment (Miller, 91, Fellbaum, et al., 1999), and this approach has now been ported to several other languages. The English lexical sample task was the result of a collaboration between Penn, which provided raining/test data for the verbs and the all-words task (Palmer, et al., 2001), and Brighton which provided the training/test data for the nouns and adjectives (Kilgarriff, 2001). Between 75 and 300 instances of each word in the lexical sample task were hand-tagged, depending on the number of senses. The data came primarily from the Penn Treebank II Wall Street Journal corpus (Marcus, et. al., 1993), but was supplemented with data from the British National Corpus whenever there were an insufficient number of Treebank instances.

3.2 Tagged corpora

For the all-words tasks, the guidelines suggest that at least 5000 words of running text be selected, and that all content words be tagged. For the lexical sample tasks, it was suggested that for each sample word, at least 75+15*n* corpus instances be chosen, where *n* is the number of senses of the word. Again, lack of resources might have precluded this much tagged data. The **Gold Standard corpus** must be replicable; the goal is to have human taggers agree at least 90% of the time. Thus, at least two human taggers were required to tag every instance of a word. Taggers are allowed to tag with multiple tags and to use special tags for proper names, and unassignable senses. See the papers in this proceedings for more details.

For the evaluation, the corpus had to be divided into a training set and a test set. The **training set** is a random subset of the Gold Standard corpus, which allows supervised systems to train. Not all tasks supplied training data, so only 'unsupervised' systems could participate (e.g., in the English all-words task – although many systems trained on other corpora such as Semcor). The **test set** is the rest of the corpus, with tags removed, on which the systems would be evaluated. It was suggested that a 2:1 ratio of training to test data be used. Although somewhat different from what is normally used in machine learning, the committee felt that having more test data would give a more realistic indication of a system's performance (since more varied contexts per word would be tested), and, moreover, unsupervised systems would be less 'short-changed'. All data sets are now in the public domain (on the SENSEVAL website).

For the English All words task, a simple baseline strategy which simply tags each head word with the first WordNet sense for the corresponding Treebank part-of-speech tag, has a score of 57% on the All words task, as compared to the best system score of 69%. For the English Lexical Sample task, the highest system scores were fined-grained: 64.2% and coarse-grained: 71.3%, compared to a baseline of 51.2%. In general the nouns and adjectives had a lower polysemy and higher scores (fine-grained, 64%, Inter-annotator agreement 85%, polysemy 4.9) than the verbs (fine-grained, 57.6%, Inter-annotator agreement, 71%, polysemy 16.28).

3.3 Sense groupings

Since some sense inventories are too fine-grained for plausible sense disambiguation, the scoring program can take into account sense hierarchies or sense groupings. Optionally, a task could provide such a grouping of senses, so that choosing any sense within the group or higher in the hierarchy would count towards a system's overall score. For example, the WordNet hierarchy was used for English nouns, whereas a separate 'grouping' was specially constructed for the English verbs (since the verbs do not have a useful hierarchy in WordNet for scoring purposes). The coarse-grained scores for the verbs with these groupings went up 10%. See the paper on the English tasks for more details.

3.4 Common data format

All tasks used a specially defined common data format for encoding the tagged and untagged corpus examples. Specifically, it accommodated the multi-lingual nature of the data by using an XML document type definition which allowed for a flexible mapping from lexical items to their textual instances. Using XML also allowed for arbitrary character encodings in the corpora. The structure was designed so that individual instances of lexical items could be associated with multiple sense tags, and allowed for discontinuous phrasal lexical items. It did not, however allow for multiple phrasal items with overlapping portions in the surface string.

Another requirement was simplicity. This quality would not only facilitate the logistics of designing a task, but would also ease any hand annotation that may have been necessary. As a result, a standoff annotation system was not feasible. A standoff annotation would restrict the format in such a way as to limit the feasibility of embedding extant annotation of the corpora and to require that participants use standoff annotation in submitting their answers for reasons of space efficiency.

The use of the common data format simplified many system' s participation in multiple tasks, consequently furthering research into the comparison of WSD in different languages.

4 Evaluation procedure

The evaluation was run centrally from a single website at the University of Pennsylvania and followed the same procedure as used in the first SENSEVAL. For each task, data was released in three stages:

- **Trial data:** A small set of data so that participants can design their systems to use the data formats. No 'real' data was released.
- **Training data.**
- **Test data.**

Each team would register their system, and then download the data sets according to the schedule. After running their system on the test data, each team submitted their answers to the website for automatic scoring. Each team's results were returned to the team before the workshop, but the overall results were unveiled at the workshop.

4.1 Schedule

A schedule was set up for task organisers to prepare and submit their data to the central website, while participants followed a separate, more rigid (and in the end very tight), schedule for downloads and submissions. Task organisers started preparing their data as far back as September 2000, but the real push occurred in the three months proceeding the competition period. The competition period ran April 17 – June 18. Within this period, each task had a critical window defined to be the period from when the training data was first made available to the last day for answer submissions to that task. The critical window had to be a minimum of 21 days. Participants could download and submit answers at any time during the critical window of a particular task, subject to the following constraints. A submission of answers must:

- not have occurred more than 7 days after downloading the test data,
- not have occurred more than 21 days after downloading the training data, and
- have occurred before the end of the critical window for the particular task

This set up allowed participants to have sufficient time to participate in several tasks over the whole competition period, while ensuring that on any particular task, a participant had a maximum of one week to run their system (and 3 weeks to train their system), which we hope did not give any time for tailoring systems to the specific words or the corpora of the competition.

4.2 Data distribution

Data for the tasks was distributed via a website at University of Pennsylvania. Participants were required to register for tasks in order to download the trial, training, and test data for the tasks, and to upload their answers. Each of these operations required authentication via a password chosen at the time of registration. Additionally, timestamps were recorded for each of these operations in order to enforce the timing constraints on a per-participant basis. The system was not secure, as a participant could register multiple times under different names and use the data from the first registration to perform the task at hand. However, there were no signs of security problems in the use of the website.

Use of the distribution center was recommended, not required, of the task organizers. All the tasks with the exception of the Japanese tasks used the distribution center. A nice by-product of this process in concert with the common data format was the development of an overarching organization of all the SENSEVAL data, which is evident in the data available to the public domain.

4.3 Scoring and evaluation

The same answer format and scoring program was used for SENSEVAL-2 that was used in the first SENSEVAL. Systems were allowed to tag a word with as many senses as appropriate, giving probabilities, if desired. If the task had a sense hierarchy or grouping, then fine- and coarse-grained scoring was 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 were not required to tag all instances of a word, or even all words, thus, as in SENSEVAL-1, we used precision and recall to score the systems, although the metrics 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 chose to give answers to; the cases might be particularly easy to disambiguate, but this can be determined by comparing the answers to the baseline on the same subset (a type of analysis that has yet to be done). **Coverage,** the percentage of instances that a system gives any answer to, is also reported. Where they are available, baseline and inter-tagger agreement numbers are given.

No further data analysis was done. Thus, the question of who 'won' depends on one's perspective, but, in fact, that is not the relevant question. The important thing is to examine how each system achieved the performance that it shows. Some of this analysis is given in the papers of this proceedings. (Note that in the results, where appropriate, we distinguished between supervised and unsupervised systems.)

When the results were unveiled at the workshop, it soon became apparent that bugs in the scoring software had potentially affected the results. It was decided by everyone present (on the first day) that all systems should be rescored. Also, owing to the tight schedule, some teams had made serious inadvertent errors in formatting their answers. Thus, it was also agreed that any team could resubmit their (corrected) answers before 31 July 2001. In so doing, the team would have to include an explanation about the modifications and only reasons of 'egregious' bugs would be allowed.

The official results list all original submissions scored with the debugged scorer, and all of the resubmissions, clearly identified. This compromise maintains the professionalism of SENSEVAL, as it does not devalue any team that met the original deadline, while encouraging the scientific purpose of the exercise.

5 Recommendations

Because the results were released so close to the workshop, there had been no time for detailed analysis. Thus, the workshop was structured around a series of panels about WSD and evaluation. Panels were held on domain-specific disambiguation, task design for new languages to SENSEVAL, sense distinctions, applications of WSD, and standardizing WordNets. Ideally, the majority of the workshop content should have been about the analysis of WSD algorithms, so the major recommendation for future exercises is to allow at least one month for analysis before the workshop. Part of this recommendation is to have a proceedings at the workshop, rather than post-workshop as this one. A related recommendation is to gather information about systems (e.g., supervised / unsupervised, knowledge source, etc.) as they are registered.

Second, the use of different granularities and groupings for the lexicons in question yielded some unnecessary inconsistency across tasks. For example, the English tasks used a

grouping which invalidated the mixed-grained scores, whereas the Swedish task used a hierarchy which yielded vacuous coarse-grained scores. This is actually a central issue in WSD, which should be addressed before the next SENSEVAL exercise. The data from SENSEVAL-2 should be invaluable in this research.

References

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

Nancy Ide and Jean Veronis (1998) Introduction to the Special issue on Word Sense Disambiguation. *Computational Linguistics Journa*l, *24(1).*

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

Adam Kilgarriff and Joseph Rosenzweig. (2000). Framework and results for English SENSEVAL. *Computers and the Humanities 34(1–2):15–48.*

M. Marcus, B. Santorini, and M. Marcinkiewicz (1993) Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19(2).*

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

George A. Miller and Christiane Fellbaum.(1991) Semantic networks of English. *Lexical and Conceptual Semantics, Cognition Special Issue, pages 197—229.*

Martha Palmer and Marc Light. (1999) Introduction to the special issue on semantic tagging. *Natural Language Engineering, 5(2).*

SENSEVAL Website: http://www.itri.bton.ac.uk/events/senseval

SENSEVAL-2 Website: www.sle.sharp.co.uk/senseval2