English SENSEVAL: Report and Results

Adam Kilgarriff
ITRI, University of Brighton
Brighton, England
adam@itri.bton.ac.uk

Joseph Rosenzweig
University of Pennsylvania
Philadelphia, USA
josephr@linc.cis.upenn.edu

Abstract

There are now many computer programs for automatically determining which sense a word is being used in. One would like to be able to say which were better, which worse, and also which words, or varieties of language, presented particular problems to which programs. In 1998 a first evaluation exercise, SENSEVAL, took place. The English component of the exercise is described, and results presented.

1 Introduction

There are now many computer programs for automatically determining which sense a word is being used in. One would like to be able to say which were better, which worse, and also which words, or varieties of language, presented particular problems to which programs. To this end, an evaluation exercise, SENSEVAL, was organised under the auspices of ACL SIGLEX (the Lexicons Special Interest Group of the Association for Computational Linguistics), EURALEX (European Association for Lexicography), ELSNET, and EU Projects SPARKLE and ECRAN). It comprised word sense disambiguation (WSD) tasks for English, French and Italian. The exercise is chronicled in a Special Issue of *Computers and the Humanities* (Kilgarriff and Palmer, 2000). In this paper we describe the structure, organisation and results of the SENSEVAL exercise for English.[1]

The form of the evaluation was as in MUC and other ARPA evaluations (Hirschman, 1998). First, all likely participants were invited to express their interest and participate in the exercise design. A timetable was worked out. A plan for selecting evaluation materials was agreed. Human annotators were set on the task of generating a set of correct answers, the 'gold standard'. The gold standard materials, without answers, were released to participants, who then had a short time to run their programs over them and return their sets of answers to the organisers. The organisers then scored the answers, and the scores were announced and discussed at a workshop.

In this paper we first outline some of the choices taken in the course of defining the task; then we describe the data that was used and the participating systems, and finally, the system results.

2 Task Design Choices

2.1 All-words *vs.* lexical-choice

Two variants of the WSD task are "all-words" and "lexical sample". In all-words, participating systems have to disambiguate all words (or all open-class words) in a set of texts. In lexical-sample, first, a sample of words is selected. Then, for each sample word, a number of corpus instances are selected. Participating systems then have to disambiguate just the sample-word instances. For SENSEVAL, the 'lexical sample' variant was chosen. The reasons included

- More efficient human tagging
- The all-words task requires access to a full dictionary. There are very few full such dictionaries available (for low or no cost)
- Many of the systems needed either sense-tagged training data or some manual input for each dictionary entry so could not have participated on the all-words task.

[1] There is a fuller version of the paper in the Special Issue. SENSEVAL materials are available at http://www.itri.bton.ac.uk/events/senseval

2.2 Dictionary and Corpus

A WSD exercise requires a dictionary, to specify the word senses which are to be disambiguated. It also requires a corpus of language data to be disambiguated. For English SENSEVAL, the HECTOR database provided both.[2]

HECTOR was a joint Oxford University Press/Digital project (Atkins, 1993) in which a database with linked dictionary and corpus was developed. For a sample of words, dictionary entries were written in tandem with sense-tagging all occurrences of the word in a 17M-word corpus (a pilot for the British National Corpus).

The primary reason for the choice was a simple one. At the time when a choice was needed, it was not evident whether there was any funding available for manual tagging. Had funding not been forthcoming, then, with the HECTOR data, it would still have been possible to run SENSEVAL as corpus instances had been manually tagged in the HECTOR project.[3]

One disadvantage of the HECTOR corpus material in the form in which it was received from OUP was that corpus instances were associated with very little context: sometimes just one sentence, by default two sentences. Strategies for gleaning information from a wider context would not show their strength.

2.3 Lexicon sampling

A criticism of earlier forays into lexical-sample WSD evaluation is that the lexical sample had been chosen according to the whim of the experimenter (or, to coincide with earlier experimenters' selections). For English SENSEVAL, a sampling frame was devised in which words were classified according to their frequency (in the BNC) and their polysemy level (in WordNet) and the sample of 35 words (corresponding to 41 tasks—see below) was then selected from the the set of HECTOR words.

[2]We are grateful to OUP for allowing us to use the HECTOR material.

[3]There was one other possible source of already-tagged data: the SEMCOR corpus, tagged according to WordNet senses (Fellbaum, 1998). However, SEMCOR was already widely used in the WSD community so SEMCOR could not provide "unseen" data for evaluation; also it adopted the all-words approach.

2.4 Word Class Issues

Word class issues complicated the task definition. The primary issue was: was the assignment of word class (POS-tagging) to be seen as part of the WSD task? In brief, the argument **for** was that, in any real application, the word sense tagging and POS-tagging will be closely related, with each potentially providing constraints to the other. The argument **against** was 'divide and rule': POS-tagging is a distinct sub-area of NLP, with its own strategies and issues, and (arguably) a high accuracy rate, so was best kept separate. A previous SIGLEX meeting had seen a majority in favour of decoupling, but no unanimity.

For English SENSEVAL, for most of the evaluation words, the tasks were decoupled, with the part-of-speech (noun, verb or adjective) of the corpus instance specified by the organisers as part of the input to the WSD task. However for five words, the tasks were not decoupled, so participating systems had to assign a sense without prior knowledge of word-class. This gave rise to a distinction between words and 'tasks'. Each SENSEVAL **task** was identified by a word and either a word-class (noun, verb or adjective) or 'indeterminate'.

3 The data

There were three data distributions. The **dry-run distribution** comprised a set of lexical entries and corresponding corpus instances and could be used to adapt systems to the format and style of data that would be used for evaluation.

The **training-data distribution** comprised the lexical entries for the test words and some sense-tagged corpus instances for most of them. The lexical entries were provided so that participants could ensure that their systems could parse and exploit the dictionary entries and add to them where necessary, and the corpus instances, so that supervised-training systems could be trained for the words in the lexical sample. For five words there was no training data, and for the remainder, the quantity varied widely between 26 and 2008 instances, depending simply on how many there were available.

In both dry-run and training data, corpus instances were provided complete with the sense-tag that had been assigned as part of the orig-

inal HECTOR tagging, but there had been no re-tagging.

The **evaluation distribution** contained, simply, a set of corpus instances for each task. Each instance had been tagged by at least three humans, though these tags were, of course, not part of the distribution. There were 8448 corpus instances in total in the evaluation data, most tasks having between 80 and 400 instances. There were 15 noun tasks, 13 verb tasks, 8 adjectives, and 5 indeterminates.

Systems were required to return, for scoring, a one-line answer for each corpus instance comprising task name, reference number and one or more sense tags, optionally with associated probabilities.

Gold standard replicability

Preparation of a gold standard worthy of the name was critical to the validity of WSD evaluation, as discussed in detail in (Gale et al., 1992). The taggings must be correct, and it can only be deemed that they are correct if different individuals or teams tagging the same instance dependably arrive at the same tag. In various manual sense-tagging exercises, agreement levels between taggers have been low. For SENSEVAL, it was critical that they were high. To this end, the individuals to do the tagging were carefully chosen: whereas other tagging exercises had mostly used students, SENSEVAL used professional lexicographers. The HECTOR dictionary was selected in part because it was corpus-based, had many examples, and was likely to support high-accuracy tagging. Taggers were encouraged to give multiple tags (one of which might be the 'unassignable' tag) rather than make hard choices. The material was multiply tagged, and an arbitration phase introduced: first, two or three lexicographers provided taggings. Then, any instances where these taggings were not identical were forwarded to a third lexicographer for arbitration.

At the time of the SENSEVAL workshop, the tagging procedure (including arbitration) had been undertaken once for each corpus instance. Individual lexicographers' initial pre-arbitration results were scored against the post-arbitration results. The scoring algorithm was as for system scores. The scores ranged between 88% to 100%, with just five out of 122 results for <lexicographer, word> pairs falling below 95%.

To determine the replicability of the whole process in a thoroughgoing way, the exercise was repeated for four words, selected to reflect the spread of difficulty. The 1057 corpus instances for the four words were tagged by two lexicographers who had not seen the data before and non-identical taggings were forwarded for arbitration. These taggings were then compared with the ones produced previously. The level of agreement was 95%. This was a most encouraging result, which showed that it was possible to organise manual tagging in a way that gave rise to high replicability, thereby validating the WSD enterprise in its entirety, and SENSEVAL in particular.

4 Systems

The seventeen systems which returned results prior to the workshop are shown in Table 1.

Systems differ greatly in terms of the input data they require and the methodology they employ. This makes comparisons particularly odious, but, to make the comparisons marginally more palatable, they were classified into two broad categories, the supervised systems, which needed sense-tagged training instances of each word they were to disambiguate, and the non-supervised systems which did not.

The scheme is a first pass, and various classifications seem anomalous. Some supervised systems are also equipped to fall back on alternative tagging strategies in the absence of an annotated training corpus, while some non-supervised systems default to a frequency-based guess if information from a training corpus is available. Systems such as SUSS and CLRES were in principle nonsupervised, but used the training data (as well as the dry-run data) to debug and improve the configuration of their programs. We use the scheme to simplify the presentation of results, but ask the reader to treat it indulgently.[4]

[4]For participants whose systems output WordNet senses, a mapping from WordNet senses to HECTOR senses was provided by the organisers. The result is not altogether satisfactory, with gaps, one-to-many and many-to-many mappings. The performance figures for the four systems (UPC-EHU-UN, UPC-EHU-SU, SUSSEX AND OTTAWA) which used the mapping suffered substantially.

Group	Shortname
Nonsupervised	
CL Research, USA	clres
Tech U Catalonia, Basque U	upc-ehu-un
U Ottawa	ottawa
U Sunderland	suss
U Sussex	sussex
U Sains Malaysia	malaysia
XRCE, CELI, Torino	xeroxceli
Supervised	
Bertin, U Avignon	avignon
Ed Testing Service, Princeton	ets-pu
John Hopkins U	hopkins
Korea U	korea
NMSU, UNC Asheville	grling-sdm
Tech U Catalonia, Basque U	upc-ehu-su
U Durham	durham
U Manitoba	manitoba-ks
U Manitoba	manitoba-dl
U Tilburg	tilburg

Table 1: Participating systems for English

Baselines

System results can be measured against two sets of baselines; one that makes use of the corpus training data, and the other that uses only dictioanry entries. The former are intended for comparison with supervised systems, the latter, for comparison with nonsupervised ones. None of the baselines draws on any form of linguistic knowledge, except for those that are coupled with the phrase filter, which recognizes inflected forms of words and applies rudimentary ordering constraints for multi-word expressions.

The highest-performing baselines were all variants of Lesk's algorithm (Lesk, 1986). The Lesk-based baselines outperformed the baselines which used simpler algorithms such as RANDOM, or, "always choose the sense which has most training-corpus instances".

Simple LESK chooses the sense of a test word's root whose dictionary definition and example texts have the most words in common with the words around the instance to be disambiguated. The strategy is, for each word to be tagged:

```
(a) For each sense s of that word,
(b)    set weight(s) to zero.
(c) Identify set of unique words W in
surrounding sentence.
(d) For each word w in W,
(e)    for each sense s,
(f)      if w occurs in the definition or
example sentences of s,
(g)        add weight(w) to weight(s).
(h) Choose sense with greatest weight(s)
```

Weight(w) is defined as the inverse document frequency (IDF) of the word w over the definitions and example sentences in the dictionary. The IDF of a word w is computed as $-log(p(w))$, where p(w) is estimated as the fraction of dictionary "documents" —definition or examples— which contain the word w.

LESK-PLUS-CORPUS is as LESK, but also considers the tagged training data, so can be compared with supervised systems. For each word in the sentence containing the test item, it tests whether w occurs in the dictionary entry or corpus instances for each candidate sense.

Although LESK-PLUS-CORPUS does not explicitly represent the relative corpus frequencies of sense tags, it favors common tags because they have larger context sets, and an arbitrary word in a test-corpus sentence is more likely to occur in the context set of a commoner training-corpus sense tag.

The baselines all performed better when coupled with a phrase filter designed to scan for multi-word expressions. It runs first, vetoing all senses for multi-word items if there is no evidence for them in the test instance, and vetoing all senses except the salient multi-word one(s) where evidence is found.

5 Results

The scoring regime allowed scores of between 0 and 1 where a system returned more than one sense for an instance, with the probability mass shared, as described in (Melamed and Resnik, 2000).[5] The precision, or performance, of a system is computed by summing the scores over all test items that the system guessed on, and dividing by the number of guessed-on items. Recall is computed by dividing the system's scores over all items by the total number of items.

[5] A number of strategies were explored for relating scores to the hierarchy of senses and subsenses in the dictionary. In this exercise, the choice of scoring scheme made little difference to the relative scores of different systems, or of systems on different tasks. In what follows, only direct sense-to-sense or subsense-to-subsense matches are considered.

The highly skewed distribution of language phenomena, with a few very frequent phenomena and a long tail of rarer ones, also that systems will primarily be evaluated with respect to their ability to handle a few common types of problems. Their ability to handle a range of rarer problems will have little impact on their score. Even if a system does not choose to restrict itself to the subset of common cases, there will be little else for it to demonstrate its versatility on.

Figure 1 summarises system performance on the overall task. Nonsupervised systems are in italics, supervised in boldface. The human score, HECTOR, corresponds to the annotations made by the lexicographers who initially marked up the test corpus.

Three baselines are also provided for comparison. (LESK does not explicitly use the corpus, but does benefit from the corpus-like dictionary examples, which are like a mini-corpus, and, for many dictionaries, would not be available. Hence the inclusion, for comparison, of LESK DEFINITIONS, which does not use this source of information.)

Figure 1 demonstrates that the state of the art, for a fine-grained WSD task where there is training data available, is at around 75%. Where there is training data available, systems that use it perform substantially better than ones that do not.

For nouns, the top performance was over 80%; for the verbs, the best systems scored around 70% with the other two categories, adjectives and indeterminates, falling in between.

The majority of systems were outperformed by the Lesk baseline for their system-type. On one large subset of the data, the 2500 items in the verb tasks, none of the systems is capable of achieving more than a 2% improvement over the best baseline's error rate.

Some of the supervised systems (`durham, hopkins, suss, manitoba-dl`) were designed to fall back on unsupervised techniques, or to rely on dictionary examples when no corpus training data was available. One might have expected these systems to perform at the same levels as nonsupervised systems for those tasks where there was no training data. But this was not the case. The supervised systems performed better even for these words.

Individual items in the dataset are not graded in any way for difficulty. This is a limitation of the evaluation since most systems did not tag the entire dataset but carved out more or less idiosyncratic subsets of it, abstaining from guessing about the remainder. Without difficulty ratings for items, we cannot say whether two systems that tag only part of the data have chosen equally hard subsets, and results may not be comparable.

5.1 Polysemy, entropy, and task difficulty

The distribution of sense tags in the training and evaluation data is highly skewed, with a few very common sense tags and a long tail of rarer ones. This suggests that the distributions of sense tags for individual words in the data will also be quite skewed and that the entropy of these distributions[6] will be fairly low. However, there is substantial variation of entropy across words. For instance, both **generous** and **slight** are adjectives with 6 senses, but the entropy of **slight** is 1.28 while that of **generous** is 2.30. This is because of the unusually even distribution of sense tags for **generous**.

Polysemy and entropy often vary together, but not always. The nouns, on average, had higher polysemy than the verbs but the verbs had higher entropy. For verbs, the corpus instances were spread across the dictionary senses more evenly than for nouns.

Systems tend to do better on the nouns than the verbs, suggesting that entropy is the better measure of the difficulty of the tasks. The correlation between task polysemy and system performance is -0.258. The correlation between entropy and system performance is stronger: -0.510. When considering just the supervised systems, the correlation with entropy is -0.699; with polysemy, -0.247.

6 Conclusion and way forward

We have presented a first open evaluation for Word Sense Disambiguation systems for English. The exercise was a success, with the various obstacles to involving different members of the community, with different varieties of WSD

[6] Entropy is calculated as $-\Sigma(p(x) \cdot log(p(x)))$ where x ranges over all sense tags of a word, and p(x) is the fraction of training occurrences of the word tagged with x.

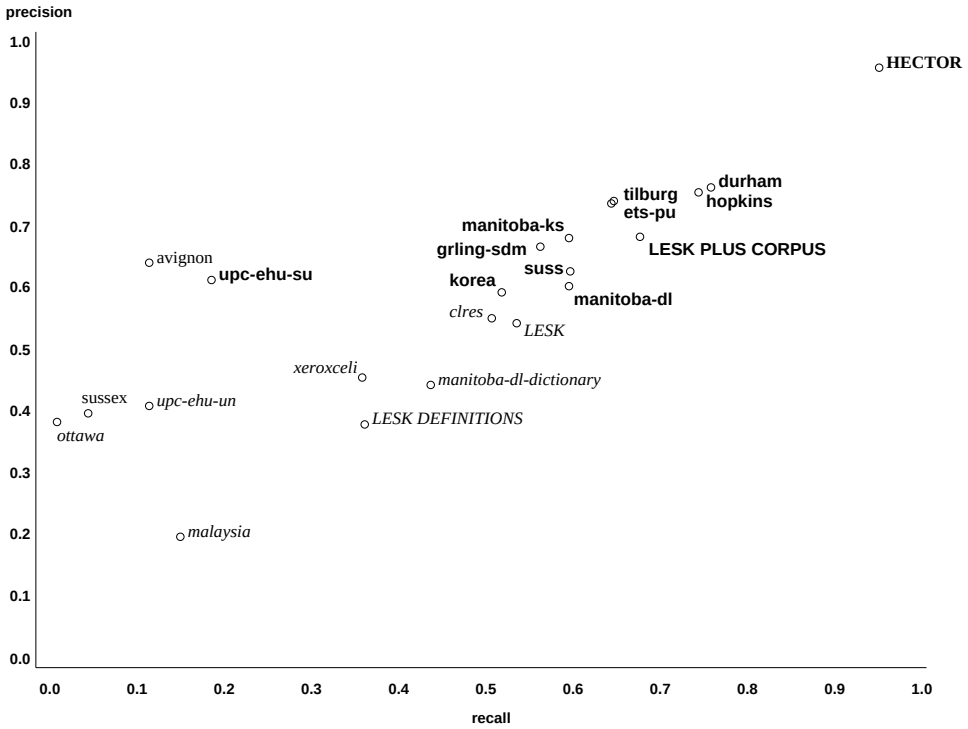

Figure 1: System performance on all test items.

system, all overcome to some degree. A notable success was the achievement of high replicability for the manually-tagged gold standard. A notable limitation was that systems which did not tag according to HECTOR senses, but according to other senses which were then mapped, were at a severe disadvantage.

The results demonstrate that the state of the art for fine-grained WSD, where there is training data available, is 75–80%. Where there is training data available, systems that use it perform substantially better than those that do not. They also demonstrate that a well-implemented simple LESK algorithm is hard to beat.

SENSEVAL demonstrates the feasibility and value of WSD evaluation exercises and we believe there should be future SENSEVALs, with the task re-designed according to the strengths and weaknesses of this first one.

References

Sue Atkins. 1993. Tools for computer-aided corpus lexicography: the Hector project. *Acta Linguistica Hungarica*, 41:5–72.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

William Gale, Kenneth Church, and David Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings, 30th ACL*, pages 249–156.

Lynette Hirschman. 1998. The Evolution of Evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language*, 12(4):281–307.

Adam Kilgarriff and Martha Palmer. 2000. Guest editors, Special Issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs. *Computers and the Humanities*.

Michael E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proc. 1986 SIGDOC Conference*, Toronto, Canada.

I. Dan Melamed and Philip Resnik. 2000. Evaluation of sense disambiguation given hierarchical tag sets. *Computers and the Humanities*, Special Issue on SENSEVAL.