SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs

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

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 this paper I describe a pilot evaluation exercise ('SENSEVAL') taking place under the auspices of ACL SIGLEX (the Lexicons Special Interest Group of the Association for Computational Linguistics) and EURALEX (European Association for Lexicography) in 1998.

1 Word Sense Disambiguation

As dictionaries tell us, most common words have more than one meaning. When a word is used in a book or in conversation, generally speaking, just one of those meanings will apply. This is not a problem for people. We are very rarely slowed down in our comprehension by the need to work out which meaning of a word applies. But it is for computers. The clearest case is in Machine Translation. If English *drug* translates into French as either *drogue* or *médicament*, then an English-French MT system needs to disambiguate *drug* if it is to make the correct translation.

For forty years now, people have been writing computer programs to do Word Sense Disambiguation (WSD). Early programs (Kelly and Stone, 1975; Small, 1980) required human experts to write sets of disambiguation rules for each multi-sense word. This was a problem. It involved a huge amount of labour to write rule-sets or "Word Experts" for a substantial amount of the vocabulary.

The WSD problem can be divided into two parts. The first is, how do you express what meaning number 1 and meaning number 2 of a word are, to the computer. The second is, how do you work out which of those meanings matches an occurrence of a word to be disambiguated. (Lesk, 1986) took a novel tack, using the text of dictionary definitions as an off-the-shelf answer to the first problem. He then measured the overlap, in terms of words-in-common, between each of the definition texts and the context of the word to be disambiguated. Much recent work uses sophisticated variants of this idea.

Dictionary-based approaches remain tied to a particular dictionary, with concomitant errors, imperfections and copyright constraints. With the advent of huge computer corpora, and computers powerful enough to computer complex functions over them, the 1990s has seen new strategies which find the contexts indicative of each sense in a training corpus, and then find the best match between those contexts and the instance of a word to be disambiguated (Yarowsky, 1995).

2 Evaluation

So there are now quite a few working WSD programs. An obvious question is, which is best? Evaluation has excited a great deal of interest across the Language Engineering world of late. Not only do we want to know which programs perform best, but also, the developers of a program want to know when modifications improve performance, and how much, and what combinations of modifications are optimal. US experience in ARPA competitive evaluations for speech recognition, information retrieval etc. has been that the focus provided by an evaluation serves to bring research communities together, forces consensus on what is critical about the field, and leads to the development of common resources, all of which then stimulates further rapid progress.

Reaping these benefits involves overcoming two major hurdles. The first is agreeing an explicit and detailed definition of the task. The second is produc-

ing a "gold standard" corpus of correct answers, so it is possible to say how much of the time a program gets it right. In relation to WSD, defining the task includes identifying the set of senses between which a programme is to disambiguate, the "sense inventory" problem. Producing a gold standard corpus is both expensive, as it requires many person-months of annotator effort, and hard because, evidence to date shows, different individuals will often assign different senses to the same word-in-context.

A workshop of the ACL Lexicon Special Interest Group (SIGLEX) in Washington, April 1997, included a lively and productive session on WSD evaluation. (Resnik and Yarowsky, 1997) made some practical proposals which were broadly welcomed. There was a high degree of consensus that the field needed evaluation, and that researchers needed to collaborate and make compromises so that an evaluation framework could be agreed.

In the subsequent discussion, there were two cultures in evidence — the computer scientists, who view a set of dictionary definitions as data they are to work with (and would like to be able to treat them as fixed) and the humanists, who had detailed experience of lexicography and textual analysis, and whose dominant concern lay in the sheer difficulty of identifying and defining word senses.

The humanists argued that a high level of agreement between different people doing the tagging was not easy to achieve because the task was hard, and existing dictionaries were not up to it. This is scarcely surprising: they were written, for the most part, to explain word meanings to people, not to make cut-and-dried distinctions between senses. But without high inter-annotator agreement, the gold standard was fool's gold. There would only be potential for high inter-annotator agreement if the dictionary and its sense inventory were of very high quality, and designed for the purpose. This could be achieved through allowing the people who were doing the tagging to improve the dictionary entry, perhaps changing the senses for the word, if they found that the corpus data they were tagging was at odds with the input dictionary (at least from an NLP perspective). They could also make much fuller dictionary entries as they would not be constrained to column inches, as paper lexicographers always are. In the Resnik-Yarowsky proposals, just 200 test words would be worked on each year, which suggested a manageable amount of lexicography-revision to undertake year on year.

Allowing shifting goalposts, in the form of a revisable sense inventory, makes for great difficulties for WSD algorithms. But to be endorsed by the research community, an evaluation framework must not only provide computable measures, but must be valid. For that, a fully defensible sense inventory and gold standard are essential.

3 Pilot SENSEVAL

The author is currently co-ordinating the first pilot WSD evaluation exercise, or SENSEVAL.

A call for participants has been published and there are over 20 systems (hereafter "the participants"), from three continents, planning to take part. Participation involves, minimally,

1. receiving corpus data from the organisers
2. applying the participant's WSD program to it
3. returning the program's word sense decisions to the organisers for evaluation.

This will take place over the summer, 1998, and there will be a workshop in Sussex, England, in September, by which time the performance of a number of WSD programs will have been evaluated, and where we shall discuss

- systems' results (from different sites, for different words etc.)
- the difficulties faced by the human lexicographers/taggers
- the way forward.

3.1 Languages covered

Most research in WSD has been on English. There are most resources available for English, most commercial interest, and most expertise in the problems in presents. ACL SIGLEX will find it easiest to set up the exercise in English. However we have no wish to be so limited, and various people working in languages other than English are involved in SENSEVAL. Ideally, there would be parallel exercises for a number of languages. By the time of the 1998 workshop, alongside the exercise for English, there will be pilots for French (5 participants), Spanish (3) and Italian (2). Preliminary planning for Korean and Portuguese is underway. Enquiries regarding setting up exercises for additional languages are most welcome.

3.2 Manually sense-tagged corpora

For English, there are various manually sense-tagged datasets in existence. Some could provide data for SENSEVAL. The survey below covers all datasets for English where a combination of size, care taken over tagging, and availability make them candidates for use in an evaluation exercise.

3.2.1 SEMCOR

The best known and most widely-used manually sense-tagged corpus is SEMCOR (Fellbaum, 1997). It comprises 250,000 words of text (taken from the Brown Corpus and a novel, "The Red Badge of Courage") in which all content words have been tagged, manually, with word sense. The sense inventory is taken from the WordNet lexical database. It is available free over the WorldWideWeb. It is a very valuable resource which has already been widely used for WSD evaluation as well as a range of other purposes, and has contributed greatly to our understanding of the task and the problems involved. One of these contributions regards the mutability of the dictionary. Originally, the plan was to be that SEMCOR taggers would not make changes to the dictionary. The SEMCOR experience demonstrated that this was not viable. Where a tagger could not make sense of a sense-distinction in WordNet, their choice of one sense over the other become arbitrary. The situation was resolved by providing an avenue for the tagger to feed into the dictionary-editing.

For SENSEVAL, SEMCOR has several shortcomings. There are only 83 words for which there are more than 100 sense-tagged corpus instances; WordNet, like any other dictionary, contains errors and inconsistencies, and these often result in anomalies in SEMCOR; and as it is freely available, it cannot provide **unseen** data for evaluation: all of SEMCOR has already been seen by many research teams in the area.

3.2.2 DSO corpus

A team in Singapore disambiguated all instances of 191 "most frequently occurring and most ambiguous" nouns and verbs in a corpus (Ng and Lee, 1996). There are 192,800 tagged tokens. Linguistics undergraduates did the tagging, and the work represents a person-year of effort. The resource is freely available and has been used by various researchers in addition to Ng and Lee.

Their data included the subset of the Brown corpus in SEMCOR, so there was some overlap between the word-instances tagged in the two projects. The level of agreement between SEMCOR and DSO taggers, with both using the full fine-grained set of WordNet senses, was 57%.

The 57% agreement with SEMCOR makes it impossible to regard the DSO corpus as a gold standard. It also indicates how hard it is likely to be to achieve a target level of 90% agreement between taggers, as is the SENSEVAL goal.

3.2.3 HECTOR

HECTOR was a joint Oxford University Press/Digital project (Atkins, 1993) in corpus lexicography. For a substantial set of words, all corpus instances in a 20M-word corpus (a pilot for the British National Corpus) were tagged according to the senses in a dictionary entry that was being developed alongside the tagging process. The database comprises 200,000 tagged instances and an associated set of dictionary entries. There are 300 words associated with over 100 corpus instances.

The tagging and the lexicography formed a single process. The tagger-lexicographers were highly skilled and experienced. There was some checking, with a second lexicographer going through the work of the first, but no extensive editing of either corpus taggings or dictionary entries. The dictionary entries are fuller than in most paper dictionaries or WordNet, and this is likely to be beneficial for SENSEVAL.

English Pilot SENSEVAL will use HECTOR data. OUP has agreed to make the data available at no cost. It will be necessary to re-tag the data to determine the level of agreement between the SENSEVAL tagger and the original HECTOR tagger.

In committing to the corpus, we are also committing, by implication, to the HECTOR dictionary entries and sense inventory. The dictionary entries are written by expert lexicographers, on the basis of particularly close scrutiny of corpus evidence — and are available electronically — so this is satisfactory.

3.3 A sample of words

SEMCOR and HECTOR represent two alternative approaches to selecting the data to be tagged. SEMCOR took the 'textual' approach, tagging everything in a selection of texts, whereas HECTOR took the 'lexical' approach, first taking a selection of word types ('dictionary headwords') and then tagging all occurrences of them in a set of texts.

Pilot SENSEVAL will prioritise the 'lexical sample' approach, for a number of reasons.

Firstly, lexical sense-tagging is not a well-understood task, and when a task is not well-understood, it is wise to find out more about it before doing a lot of it. SENSEVAL needs to assess how to learn most from limited manual sense-tagging resources. Very little can be inferred where a person sense-tags less than twenty or thirty instances of a word: there is simply insufficient evidence to draw any conclusions. In the textual approach, much of the tagger's effort is spent on word-types for which less than twenty tokens get tagged. In the lexical approach, we can decide will have at least fifty tokens

per type will be tagged.

Secondly, taggers can tag more efficiently and accurately if they work lexically rather than textually. Experience of tagging is commonly that the bulk of the intellectual labour goes into the close reading of the dictionary definitions: only when they are fully and clearly understood can non-obvious tagging decisions be made (Kilgarriff, 1993). In the lexical approach, one close reading of a dictionary entry serves for tagging a substantial set of occurrences for that word. The textual approach is inefficient, because, for each word, the tagger must look closely at a new dictionary entry. The lexical method also promotes the use of patterns. When a tagger notices a recurring pattern in the corpus lines for a word, they are usually able to infer that that pattern always signifies a particular sense. A good tagging methodology will promote the use of patterns, as was done in HECTOR.

Thirdly, with the lexical approach, SENSEVAL will only be considering a small number of word types. It will be necessary to manually establish mappings between one dictionary's senses and another's, so this will be a manageable task. Also the copyright and data-handling issues relating to whole dictionaries are avoided.

Fourthly, one class of WSD systems is only able to tag a set of word-types for which there has been some specific input. Their participation would be severely limited if a textual approach was adopted. (Developers of systems which disambiguate all words may claim to be at a disadvantage when compared with systems requiring specific input for the test-words: however the situation is not symmetrical because 'all-words' systems can participate fully in a 'lexical-sample' evaluation. See also next section.)

For pilot SENSEVAL, around 60 word-types are being selected for each language, covering nouns, verbs, and adjectives. Between 50 and 300 instances of each word-type will be manually tagged. For the approach adopted for French and Italian see (Véronis et al., 1998). For English, the choice of words is constrained by the HECTOR data: all the words in the sample must have HECTOR lexical entries and over 100 tagged HECTOR instance. The sample will cover, as far as possible, higher and lower frequency words and higher and lower polysemy words. (For a fuller discussion of sampling issues, see Kilgarriff, 1997.)

We anticipate that the difference in scale between a 'pilot' and a full-scale SENSEVAL will primarily be a difference in the number of sample words. (The pilot may well reveal more profound ways in which the model needs to change.)

3.4 Level playing fields and the Himalayas

The WSD systems involved in the English task represent a great variety of algorithms and approaches. Some rely heavily on dictionaries, others do not use dictionaries at all. Some sense-tag all words; others, only nouns, or only verbs, or only nouns occurring as heads of direct objects noun phrases; others again require some particular preparatory work for each word type to be tagged. Almost half the systems use "supervised training" methods: they require a set of sense-tagged "training" data in order to learn how to tag further examples.

The rallying cry for evaluation exercises is that there should be a level playing field. The way in which the exercise is set up and administered should not favour one participant over another. This is our goal. However, where different systems require such radically different inputs, it is not easily achieved. In practice, in all likelihood **every** participating system will be presenting its results with qualifications: the performance was not what it might have been because the dictionary had a radically different format, or because a WordNet-based semantic hierarchy could not be used, or because longer documents had been expected ... Would-be participants may well be deterred from participating by the fear that their system appears to perform badly, owing to a mismatch between the evaluation setup and their system.

The organisers' response is that, firstly, we shall level the playing field as far as we are able; secondly, we shall strongly discourage citation of results **without** reference to qualifying factors; and thirdly, the widely-shared premise is that the whole field stands to gain from evaluation. We ask would-be participants to weigh the longer-term benefits of participation, both for themselves and for the community at large, against the possible short-term embarrassment.

The HECTOR database has the merit that it has not been used for WSD research before, so no system has 'home advantage'.

Systems will not be required to use the HECTOR sense inventory directly, but for those that do not, the research group will have to produce a mapping from their sense inventory to HECTOR's.

4 The difficulty of manual tagging

Pike can mean 'fish' or 'medieval weapon':[1]

> The carp and pike, which were found locally, were kitted out with lavish trim-

[1] All citations from the British National Corpus.

mings and served ...

Towards the close of the twelfth century the pike was used to counter cavalry charges, ...

All citations from the British National Corpus.

The manual tagger's task is to say, for each of the corpus instances, whether the word is being used in its 'fish' or 'medieval weapon' sense (or neither or both). In general, the tagger first looks at a dictionary, to find what senses the word has, and then at the context, to see which sense applies. For most instances of most words, given a small context of two or three words preceding and following the target word, it is immediately apparent which sense holds. However for many words, the distinctions are not as clear cut as for *pike*, and for many instances, the selection of the appropriate sense will not be effortless. *Application* can mean, amongst other things, the document or the process of applying for something: it requires close reading to determine which applies in the following cases.

Application for a grant should be made at the same time as the application for an audition ...

I then found my application for financial assistance for part-time study had been rejected ...

The scale and difficulty of the task depend very substantially on how many words are like *pike*, and how many like *application* (and how many of the instances of each are 'straightforward'). There is almost no previous research on this point, and nothing that approaches the sampling question systematically. The SENSEVAL pilot will gather data and experience. From a practical point of view, this will support more accurate budgeting for future SENSEVALS. From a theoretical one, it will shed light on a central question about the lexicon: how, how often, and in what ways, are words used in ways that deviate from their staple meanings.

Website

`http:www.itri.bton.ac.uk/events/senseval`

Acknowledgements

I would like to thank Oxford University Press, Cambridge University Press and Addison Wesley Longman for their material support for SENSEVAL. The work was also supported by EPSRC grant GR K18931.

References

Sue Atkins. 1993. Tools for computer-aided lexicography: the Hector project. In *Papers in Computational Lexicography: COMPLEX '93*, Budapest.

Christiane Fellbaum, editor. 1997. *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press, Cambridge, Mass. forthcoming.

Edward Kelly and Philip Stone. 1975. *Computer Recognition of English Word Senses*. North-Holland, Amsterdam.

Adam Kilgarriff. 1993. Dictionary word sense distinctions: An enquiry into their nature. *Computers and the Humanities*, 26(1–2):365–387.

Adam Kilgarriff. 1997. Sample the lexicon. Technical Report ITRI-97-01, ITRI, University of Brighton. `http://www.itri.bton.ac.uk/techreports`.

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.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *ACL Proceedings*, June.

Philip Resnik and David Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In Marc Light, editor, *Tagging Text with Lexical Semantics: Why, What and How?*, pages 79–86, Washington, April. SIGLEX (Lexicon Special Interest Group) of the ACL.

Steven L. Small. 1980. *Word Expert Parsing: A Theory of Distributed Word-Based Natural Language Understanding*. Ph.D. thesis, Department of Computer Science, University of Maryland, Maryland.

Jean Véronis, Valérie Houitte, and Corinne Jean. 1998. Methodology for the construction of test material for the evaluation of word sense disambiguation systems. In *2nd Workshop on Lexical Semantics Systems*, Pisa, April.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivalling supervised methods. In *ACL 95*, pages 189–196, MIT.