University of Brighton

ITRI-98-08

Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs

Adam Kilgarriff

August, 1998

Also published in Computer Speech and Language 12 (3)

The work was supported by the EPSRC under grant K18931

Information Technology Research Institute Technical Report Series

ITRI, Univ. of Brighton, Lewes Road, Brighton BN2 4GJ, UK
TEL: +44 1273 642900 EMAIL: `firstname.lastname@itri.brighton.ac.uk`
FAX: +44 1273 642908 NET: `http://www.itri.brighton.ac.uk`

Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs

Adam Kilgarriff
ITRI, University of Brighton, Lewes Rd., Brighton BN2 4GJ, England

Abstract

There are now many computer programs for automatically determining the sense in which a word is being used. One would like to be able to say which are better, which worse, and also which words, or varieties of language, present particular problems to which algorithms. An evaluation exercise is required, and such an exercise requires a 'gold standard' dataset of correct answers. Producing this proves to be a difficult and challenging task. In this paper I discuss the background, challenges and strategies, and present a detailed methodology for ensuring that the gold standard is not fool's gold.

1 Introduction

There are now many computer programs for automatically determining the sense in which a word is being used. One would like to be able to say which are better, which worse, and also which words, or varieties of language, present particular problems to which algorithms. An evaluation exercise is required. A pilot ('SENSEVAL') is taking place under the auspices of ACL SIGLEX (the Lexicons Special Interest Group of the Association for Computational Linguistics) and EURALEX (the European Association for Lexicography) in 1998. The essential elements of a quantitative evaluation exercise are a task definition, a 'gold standard' dataset of correct answers to evaluate against, and a framework for administering the evaluation with the requisite credibility and accountability to the research community. This paper addresses the production of the gold standard.

Human tagging is an expensive, labour-intensive process and it is appealing to reuse existing manually-tagged resources. Section 3 reviews all that are available.

The pervasive worry in preparing the dataset is that it will not meet adequate standards of replicability: that is, if two people tag the same text, they will all too frequently assign different tags to the same corpus instance. The central argument of the paper is that this is a far-reaching and difficult topic, and that a high degree of replicability can be achieved, but only if the dictionary that provides the sense inventory and the individuals doing the tagging are chosen with care. Sections 4, 5 and 6 all address the issue of replicability, from various angles.

Another core issue is how the corpus instances to be tagged should be selected: the sampling question. There are two models in existence, 'textual', where all content words in a given text are tagged, and 'lexical', where first, a set of words is sampled from the lexicon, and then, a set of corpus instances for each of those is tagged. Section 7 argues the case for the lexical approach, and provides an implementation in some detail.

Most research in WSD looks at English. Most resources, commercial interest, and expertise regarding the problem it presents are tied to English. There is most momentum to set up the exercise for English. However the WSD community has no desire to be narrowly monolingual. There is nothing specific to English in the task, and various people working in languages other than English are involved in SENSEVAL. Currently, languages for which pilot SENSEVAL will run are English, French and Italian.

Terminology: types and tokens, morphology

The type-token distinction is critical to the discussion below. The word *word* is ambiguous between the two: thus there are either two, or three, words in the sentence "Dog eats dog" depending on whether one counts types or tokens. In this paper, I use *word* or *word-type* for the 'type' reading, and *token* or *instance* for the 'token' reading, which is always embedded in a particular linguistic context.

Throughout, I view word-types as lemmas. That is, in

John loves Mary.
Fred was loved by Doris.
Xavier will love Yolande.

there are three tokens of the single verbal type, *love*. Lemmatisation and part-of-speech tagging will clearly interact with WSD, though in this paper they are not discussed.

2 Background

Open a dictionary at random, choose a word at random – the odds are, the dictionary says it has more than one meaning. When a word is used in a book or in conversation, just one of those meanings will usually apply. For people this does not present a problem. Communication is very rarely impeded by the need to work out which meaning of a word applies. But for computers it is a great problem. 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 an analysis of the role of word sense disambiguation in relation to different varieties of language-engineering application, see Kilgarriff (1997).

People use the surrounding context to select the appropriate meaning. The context can be grammatical (if modified by a proper name, as in "AIDS drug", it is probably *médicament*), or lexical (if followed by *addict*, *trafficker* or *squad* it is *drogue*), or domain-based (if the text is about policing, probably *drogue*, if about disease, probably *médicament*). People can usually disambiguate on the basis of very little surrounding context, with five words generally proving sufficient (Choueka and Lusignan, 1985).

2.1 What is a word sense?

Discussions of word sense disambiguation tend to talk about 'word senses' as if they were unproblematic basic objects. But this is far from the case: preliminary evidence is that different dictionaries very often give different sets of senses for the same word, and further evidence comes from problems with the sense tagging task for humans, and the lack of operational criteria for determining where one sense ends and another begins. While these themes are never far from the topic of the paper, they are not the topic, and here we merely note that the question, “what is a word sense?”, has a long history and no simple answer.

2.2 Word Sense Disambiguation Programs

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 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. First, how do you express what meaning 1 and meaning 2 of a word are, in a way that a disambiguation algorithm can interpret. Second, how do you work out which of those meanings matches an instance of a word to be disambiguated. Lesk (1986) took a novel tack, using the words in 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 (Wilks, Slator, and Guthrie, 1996; Cowie, Guthrie, and Guthrie, 1992; Rigau, Atserias, and Agirre, 1997).

With the advent of huge computer corpora, and computers powerful enough to compute complex functions over them, the 1990s has seen new strategies which, first, find the contexts indicative of each sense in a training corpus, and then, identify the best match between those contexts and the instance of a word to be disambiguated. Some of these methods also use lexical resources (Dagan and Itai, 1994; Yarowsky, 1992; Lin, 1997; Karov and Edelman, 1998). Others, for reasons that include not wanting to be tied to a particular dictionary, errors and imperfections, copyright constraints, and lack of specificity to a particular domain, do without dictionaries (Clear, 1994; Yarowsky, 1995; Schütze, 1998). See Ide and Véronis (1998) for a recent survey of the field.

2.3 Evaluation

There are now numerous working WSD programs. An obvious question is, which is best? As witnessed by this Special Issue, evaluation is a theme of great interest to the Language Engineering world. Researchers, funders and users would all like to know which programs perform best. Developers of a program want to know when modifications improve performance, and how much, and what combinations of modifications are optimal. US experience in competitive evaluations for speech recognition, information retrieval and information extraction has been that the focus provided by a well-designed quantitative evaluation exercise serves to bring research communities together, identifies the most successful techniques, 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 (Hirschman, 1998; asnd Lin Chase, 1998).

SENSEVAL adopts the US model. The bare bones of the evaluation are:

1. definition of the task

2. selecting the data to be used for evaluation

3. production of correct answers for the evaluation data

4. distribution of the data to participants in the evaluation
5. participants use their program to tag the data, and return their taggings to the administrators
6. administrators score the participants' taggings against the gold standard
7. participants and administrators meet to compare notes, learn lessons, and to work out how future evaluations should proceed.

For Word Sense Disambiguation (WSD) evaluation, 'task definition' and 'gold standard production' are intimately linked, as they both involve the vexed question of determining a sense inventory, as discussed in detail below.

2.4 Contrast with part-of-speech tagging

The case for sense-tagging is usually developed by analogy to part-of-speech (POS) tagging and its successes. However, while there are some similarities, there are also marked contrasts. Syntactic tags such as NOUN, VERB, NP etc. are uncontentious, and the definitions of the categories have been refined by grammarians over the years. For sense-tagging, there are no such general categories. Each word type presents a new and different sense set. The authority for any particular sense set rests with a particular edition of a particular dictionary.

A new tag set implies a new disambiguation task. While POS-tagging is one task, WSD is as many tasks as there are ambiguous words in the lexicon.

For most Language Engineering purposes, the primary goal of POS-tagging is as a preliminary to parsing. This gives a focus to POS-tagging: it should provide classifications a parser can use. For sense-tagging, there is no single dominant purpose to which the tags will be put. Motivations include lexicography, information retrieval, lexical acquisition, parsing, information extraction and machine translation. It is likely that different sets of senses will be salient for each task, and for translation, a different set of senses is required for each language pair. This makes sense-tagging much the harder task, or cluster of tasks, to define.

POS-tagging is clearly a useful point of comparison for word sense tagging. However care should be taken before overworking the analogy.

3 Resources

3.1 Lexical resources

No dictionary is perfect and no two dictionaries agree on the sense inventory of a language, yet SENSEVAL cannot proceed without a sense inventory for each language it works on. The criteria used for selecting dictionaries were largely pragmatic:

1. availability in electronic form
2. no legal complications
3. quality

4. neutrality: it should not favour one group over another (because, eg, one group had been involved in its creation)
5. availability of associated manually sense-tagged data (relevant only for English)

For French and Italian, small commercial dictionaries available on CD-ROM were selected. For English, the candidates given extensive consideration were HECTOR (Atkins, 1993) and WordNet (Miller, 1990), for reasons related to 5 above.

3.2 Existing manually sense-tagged datasets

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 WordNet senses. 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 becomes arbitrary. The situation was resolved by providing an avenue for the tagger to feed into the dictionary-editing. In KILO, a follow-up project, tagging and dictionary revision are closely interwoven.

Another contribution was confirmation of the combined difficulty and tedium of tagging.

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%.

While the resource is superficially of a suitable design for SENSEVAL, the 57% agreement with SEMCOR makes it impossible to regard the DSO corpus as a gold standard.

3.2.3 Cambridge University Press

Harley and Glennon (1997) report on a sense tagger in use at Cambridge University Press, and its evaluation. The evaluation used 4000 tokens which had been hand-tagged according to the nested, very-fine-grained sense inventory of the Cambridge International Dictionary of English (CIDE, 1995), by the lexicographers working on the dictionary. The sample to be disambiguated was selected on a sentence-by-sentence, rather than word-type-by-word-type basis. The data is available for research.

3.2.4 Bruce, Wiebe *et al.*

Bruce and Wiebe (1994), Wiebe et al. (1997) and (Bruce and Wiebe, 1998) report on a series of exercises in manual tagging, explicitly within the context of WSD training and testing. In the first exercise, 2369 sentences containing the noun *interest* (or its plural form, *interests*) were tagged. More recently, a total of 6,197 tokens of 25 very high-frequency verbs were added. "How distinguishable the senses are from one another" (Wiebe et al., 1997, p 8) played a role in the selection of verbs. Work on nouns and adjectives is currently under way, with words being chosen on the basis of co-occurrence with the verbs already tagged, so mutual disambiguation (as in Hirst (1987)) can be explored. All tagging was according to WordNet senses in the first instance, though additional, finer-grained classifications have been introduced where specific uses can be accurately identified by syntactic and lexical criteria. All data is being placed in the public domain.

3.2.5 HECTOR

HECTOR was a joint Oxford University Press/Digital project (Atkins, 1993). The sample of word-types comprised the ca. 300 word types having between 300 and 1000 occurrences in a 17M-word corpus (a pilot for the British National Corpus [1]). For each of these, all corpus instances were tagged according to the senses in a dictionary entry that was being developed alongside the tagging process. Thus the tagging and the lexicography formed a single process. The tagger-lexicographers were highly skilled and experienced. There was some editing, with a second lexicographer going through the work of the first, but no extensive consistency checking.

The resource has been made available for SENSEVAL under licence from Oxford University Press. The dictionary entries are fuller than in most paper dictionaries or WordNet, and this is likely to be beneficial for SENSEVAL. They have not been edited to the degree that entries in a published product are.

3.3 Resources: Conclusion

For French and Italian, there were no manually sense-tagged corpora available, so the choice of sense inventory was not constrained by reference to corpora.

For English, there are various manually sense-tagged corpora, most of which are tagged according to WordNet senses.

WordNet has the great merits, from the research community's perspective, of being free, without licensing constraints, and available by `ftp`.[2] It also approaches the status

[1] See `http://info.ax.ac.uk/bnc`

[2] WordNet is described as a lexical database rather than a dictionary. From the point of view of dividing a word's meaning into senses, it is, however, equivalent to a dictionary and we treat it here as another dictionary.

of a *de facto* standard, as so much WSD and other NLP research has used it. WordNet versions for several other languages are currently under development in the EuroWordNet project and elsewhere (Vossen et al., 1997). Use of WordNet and EuroWordNets opens up the prospect of cross-linguistic exploration of polysemy and WSD using matched lexical resources.

It is likely that WordNet and EuroWordNets will play a central role in future evaluations. However, for pilot SENSEVAL, there was no manually sense-tagged data, tagged according to WordNet senses, which had not yet been made publicly available. Also, arguably, those systems which were designed to use WordNet had an unfair advantage over other systems.

The HECTOR corpus is, by contrast, unseen by the WSD community, and no WSD systems are specifically designed to use it. Its sense definitions, being fuller and being tailored to the corpus it comes with, are likely to facilitate the hand-tagging task. The HECTOR data is being used in SENSEVAL. It is being re-tagged, according to the HECTOR sense inventory, to determine the level of inter-tagger agreement (ITA).

4 Previous work on WSD evaluation

Gale, Church, and Yarowsky (1992) (GCY) present an extensive discussion of the WSD evaluation problem. They review earlier WSD work and note that some words are hard for WSD programs, others easy, and, to assess how effective the programs are across language in general, a random sample is required.

They provide a table where, for each of twelve words, they present performance figures for Yarowsky's (1992) system and one or more other WSD systems from the literature. Yarowsky's system, with an average score of 92%, is clearly most impressive. However GCY warn us to approach the table with caution:

> ...there are many potentially important differences including different corpora, different words, different judges, differences in treatment of precision and recall, and differences in the use of tools such as parsers and part of speech taggers, etc. (p 252)

They then consider the 'upper and lower bounds' for a WSD system. The upper bound is defined by the amount of time that people agree on the sense to be assigned, and this issue is taken up in section 6. The lower bound is the performance that a naive algorithm could achieve: they propose a model in which, first, the most common sense for a word is established (by whatever means), and then the lower bound is defined as the percentage correct if all words are assigned to the most frequent sense. For the same twelve words considered earlier, they show that the baseline varies between 48% and 96%. For some words, exceeding the baseline will not be easy.

Much recent work has had excellent results. I mention just a sample. Wilks and Stevenson (1997), tagging all words in a text with a dictionary-based algorithm, achieved 86% correct homograph assignment. Harley and Glennon (1997) used the CIDE (1995) sense inventory, which has a 'coarse' level (average 3 senses per word) and a 'fine' level (average 19). They report 78% correctness at the coarse level, and 73% at the fine level. Yarowsky (1995) has an average success of over 96% when he provides a small amount of human input in relation to each word.

The task is easier if only coarse-grained distinctions are considered, and this is the tack taken in all the work cited here, bar Harley and Glennon's second result. One researcher's coarse-grained distinction does not match another's. Some researchers have developed their own small manually disambiguated test sets. The authors using the WordNet sense inventory have been able to evaluate using SEMCOR, the DSO corpus, or Wiebe et al.'s data.

Work has followed one of two tacks: either a 'token-by-token' approach (Wilks and Stevenson; Harley and Glennon), in which case results are for all word-types taken together, or a type-by-type approach, in which results are given type by type.

5 Resnik and Yarowsky Proposals

At an ACL SIGLEX workshop in Washington in April, 1997 (Light, 1997), Resnik and Yarowsky prompted an extensive discussion on WSD evaluation, which in turn led to SENSEVAL. Here I first summarise their paper, then the discussion that followed.

Resnik and Yarowsky first make the observations that:

1. WSD evaluation is far from standardised;
2. Different tasks bear different relations to WSD, so, eg., information retrieval may fare best with a quite different approach to WSD to that required for machine translation;
3. Adequately large sense-tagged data sets are hard to obtain;
4. The field has only just begun to narrow down approaches, and identify which ones work well and which do not.

They then made four proposals:

1. **Evaluation criterion:** Current forays into WSD evaluation mostly allow only exact hits, scoring 1, or anything else, scoring 0. An alternative scheme would give a positive score to any reduction in the level of ambiguity, so a program which rejected the 'laundry' sense of *iron*, but did not choose between the 'golf' and 'metal' ones (one of which was correct) would get a positive score of less than one.
2. **Minor errors and gross errors:** If *bank* means sand bank in a sentence, then a WSD programme returning the river bank sense is doing better than one returning money-bank. The evaluation metric should reflect this, again assigning a positive score of less than one.
3. **A framework for common evaluation and test set generation:** This was their detailed proposal about how the community should set about producing a gold standard corpus.

   Each year, a fresh subset of a huge corpus is used; one part of this is reserved for hand-tagging for evaluation, and the remainder, released for training. A sample of, say, 200 ambiguous words (types not tokens) is then chosen to be used for evaluation. Each instance of each of those words in the evaluation subcorpus is manually tagged. The community does not discover what the words are until their software is frozen for evaluation, so there is no risk of the software being optimised for those particular words. A new sample of test-words is selected each year.

A major concern was that both supervised and unsupervised learning algorithms should be able to use the same evaluation corpus. To this end, any gold-standard, tagged material should be made available as training data to researchers exploring supervised learning methods as soon as this was possible without compromising the "unseen" nature of the evaluation corpus. This suggests an annual cycle in which last year's test data becomes next year's training data, with algorithms requiring training data being evaluated, each year, on last year's sample of word-types.

4. **A multilingual sense inventory for evaluation:** This was a bid to address the fraught issue of sense inventories. The aim was to apply the principle that if a word had two meanings sufficiently different to receive different translations, then the meanings were treated as distinct senses.

The working session broadly welcomed the proposals. All were concerned to develop a plan that was workable, both technically and politically, rather than one with theoretical credentials. It would need to command widespread support in the community and to be likely to attract funding. Evaluation will only count as a success if all or most actors approve the method and accept the results.

In the course of the discussion, it became apparent that the central difficulties lay in reaching a consensus between two cultures: 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, textual analysis and similar, and whose dominant concern lay in the sheer difficulty of identifying and defining word senses.

The thesis came from Resnik and Yarowsky, in the computer scientists' camp. The antithesis was that it was hard to get high inter-tagger agreement. Without that, the gold standard would be fool's gold.

A symptomatic issue was concerned multiple taggings: should a human tagger be allowed to say that more than one sense of a word applies to a corpus instance of the word (so there are multiple correct answers in the gold standard corpus)?

The computer scientists were initially unenthusiastic, since it makes the gold standard harder to use, and performance statistics more complex to define and interpret. But the humanists were adamant that sometimes, multiple correct answers were simply the truth of the matter, and the cost of defining this possibility away was that the gold standard would contain untruths. The computer scientists then started considering more sophisticated evaluation measures, which could provide scoring schemes for multiple correct answers. (The question relates closely to grain-size, and nesting of senses, since a more specific and a more general sense are often both valid for a corpus instance). The matter went to the vote and it was agreed that multiple correct answers should be retained as a possibility, though the human taggers should be discouraged from giving multiple answers unless they were clear that a single answer would be untrue.

6 The Quest for High ITA

A gold standard corpus is only worthy of the name if the answers it contains are indeed correct. Evidence to date suggests that people often disagree on the sense to be assigned to a corpus instance of a word. Studies, all in relation to WordNet, are reported in Jorgensen (1990), Fellbaum et al. (1996) and Bruce and Wiebe (1998). Jorgensen found an average agreement level on the appropriate sense for a corpus instance of just 68%. Fellbaum et al found that 'naive taggers' agreed with experts 74% of the time on average. Once Bruce et al. had eliminated their anomalous tagger, they achieved a κ score of 0.898.[3] Clearly, there is nothing trivial about obtaining a set of correct answers from humans.

The production of a gold standard corpus must therefore be approached through:

- requiring more than one person to assign senses (or 'tags')
- calculating inter-tagger agreement (ITA)
- determining whether ITA is high enough.

If ITA is not high enough, then we do not have a gold standard corpus.

ITA defines the upper bound for how well a computer program can perform. If a second human agrees with a first only 80% of the time, then it is not clear what it means to say that a program was more than 80% accurate.

Where ITA is not high, something must be done (beside changing career). GCY respond by changing the task from one of classifying word-instances according to the sense in a lexicon, to one of simply saying whether two corpus instances exhibited the same meaning of the word or not:

> Of course, it is a fairly major step to redefine the problem from a classification task to a discrimination one, as we are proposing. One might have preferred not to do so, but we simply don't know how one could establish enough dynamic range in that way to show any interesting differences. (p. 254)

(The 'dynamic range' they refer to is the range between a lower bound, which a naive WSD program could achieve, and the ITA as upper bound. They establish a lower bound which is substantially **higher** than the 68% ITA, taken from Jorgensen (1990) – hence the problem.)

The tactic is highly problematic. If the task that NLP systems need to perform is the classification one, then we have lost any clear relation between scoring on the GCY task, and ability to perform the task we wish to perform. Also, given such a low ITA, there seems little reason to proceed with the enterprise at all since we have such a weak grasp on what it means to do the task successfully.

Fortunately there is another response to low ITA: raise it! Taggings fail to agree for one of three reasons: because of an irreducible indeterminacy or ambiguity in the data, because the tags were poorly defined, or because one or more of the individuals made a mistake. All three of these can be addressed – even the first. The first can be addressed by providing a tagging scheme which allows taggers to state that an instance is ambiguous or indeterminate between two tags (or that both apply simultaneously, or that none apply). The second can be addressed by making the tags – for WSD, the sense definitions – clearer and more explicit. The third, by ensuring that the taggers are experts in the field, who fully understand the issues at hand and the distinctions to be made, and are motivated to

[3] There are a number of statistics for calculating ITA. One can take the total number of actual agreements between taggers and divide by the total number of possible agreements, but this does not allow for the number of agreements one would expect by chance. A better measure is κ (Krippendorf, 1980; Carletta, 1996), which compensates for the number of options the taggers had to choose between, and hence the level of agreement expected by chance. For current purposes, it is sufficient to note that whichever statistic is chosen, 1 (or 100%) represents complete agreement, and the greater the deviation from 1, the lower the level of agreement.

think carefully and accurately before making judgements. The second and third considerations are interrelated. Unless motivated experts are doing the tagging, there is little point in producing a sophisticated tagging scheme or lengthy tag or sense definitions. Without adequate background and motivation, the individual will be unable to keep all the distinctions in mind or apply them judiciously.

Samuelsson and Voutilainen (1997) provide a striking example of the efficacy of this approach in relation to POS-tagging. By way of background: an earlier assessment of the EngCG POS-tagger had found that it gave correct analyses to 99.7% of words. However Church (1992) presented an upper bound of 97%, which, if valid, would render the 99.7% claim meaningless. To defend the claims made for EngCG, Samuelsson and Voutilainen needed a gold standard corpus with ITA above 99.7%.

Two linguists, both expert in the EngCG framework, each tagged a 55,000-word corpus, referring to the extensive and detailed documentation wherever necessary. First they worked independently, and this gave 99.3% identical analyses. They then examined points where they had disagreed:

> virtually all were agreed to be due to clerical mistakes. Only in the analysis of 21 words, different (meaning-level) interpretations persisted, and even here both judges agreed the ambiguity to be genuine. (p 247)

Thus Church's 97% upper bound is shown to be overly pessimistic. The discouraging conclusion that there is a margin for which POS-tagging is simply not a well-defined task is disproved. It was a consequence of an insufficiently sophisticated linguistic framework, possibly combined with other imperfections in the tagging procedure. It is not critical to the argument that the agreement rate only exceeded 99.7% after taggers' results were compared. The object of the exercise is to achieve a correct set of taggings, and for this, it is sufficient that the experts agree and that the result is reproducible. An experimental setup in which taggers work independently is an important means to that end, but is not in itself critical.

100% ITA is an unrealistic goal for SENSEVAL, but Kappa of over 0.8 (equivalent to between 80 and 90% raw ITA, depending on the number of senses per word) is critical to the viability of the exercise. To this end, great care will be taken over tagging frameworks, lexical resources — and individuals.

6.1 Dictionary improvement

The sense-tagging task is the mirror of the lexicographer's task. The lexicographer takes corpus instances of a word and puts them into separate heaps, calls each heap a distinct word sense, and writes a definition for it. The tagger is given the definitions for each heap and allocates corpus instances to them. The validity of each task is constrained by the validity of the other.

A human can only tag consistently and coherently if the dictionary that provides the sense inventory is well-written, makes sense distinctions intelligently and clearly, and provides a reasonably full specification of the senses, which is usually best done by providing several examples. It must have well-defined policies on, *inter alia*,

- collocations and multi-word expressions
- nesting of senses;

- regular polysemy;
- partially-conventionalised metaphor and metonymy.

If the computer scientists' bottom line for the SENSEVAL task is that scoring had to be possible, the humanists' is that the lexicography must not be immutable. Whatever dictionary is used as a starting point for a manual sense-tagging exercise, it will not be perfect, and its imperfections will be snags upon which the project will founder. Where the taggers find it is impossible to tag the corpus instances for a word accurately, because the dictionary entry they are working to is wrong, or vague, or incomprehensible, then it has to be possible for them to revise it.

There are various forms the revision might take. The least problematic is the addition of further examples and other information to amplify a given sense and clarify where its boundaries lie with respect to other senses.

Then there is the introduction of more structure to entries. There was widespread agreement at the Washington meeting that nested entries were desirable. Without them, it would not be possible to give more credit to near misses than gross errors, or to use the same data set to assess both coarse-grained and fine-grained disambiguation. Most lexical resources do not straightforwardly encode any but the most rudimentary nesting of senses.

Then there is the addition of new sense entries. Any exercise in corpus lexicography throws up collocations which are idiomatic to some degree and which will not already be in the dictionary so are potential additions to the entry.

Most problematic of all will be those cases where the tagger finds that the analysis of the word's meaning into senses is incompatible with the corpus evidence for the word, so a new analysis is required before tagging can proceed in a principled manner. (See Stock (1983) on *culture* for an example of plausible yet incompatible analyses.) It is to be hoped that such cases are rare.

The mutability of the dictionary will have repercussions for several aspects of the Resnik-Yarowsky proposals. The final state of the dictionary will not be fixed or made public until a short while before the evaluation. This will create difficulties for those training strategies which depend on substantial pre-compilation of the dictionary. Moreover, the segment of the lexicon to be used for the evaluation will be systematically different from the remainder: it will have more extensive entries, with more examples and more nesting.

The original proposal argued that the test-set of word-types should be unseen. It is unlikely that this could be kept to in its entirety, since, as soon as the 'frozen' dictionary was made public, any user could establish that those words with changed definitions were likely to be in the test set. Rather, two dates are salient: one where the test-set of words is announced and their dictionary entries published, and a later one when the evaluation set of corpus instances is distributed.

6.2 Model tagging procedure

All target words should be tagged by at least two independent taggers (as is standard practice in US ARPA evaluations). ITA should be tracked at all points. Built into the tagging process will be the possibility of correcting and adding to any inadequate dictionary entries encountered. This is, therefore, an exercise requiring experts.

The model procedure has three steps; in the first, the two experts survey the corpus data and determine what, if any, changes to the dictionary entry are required, first independently and then together. In the second, they will tag independently. In the third, they will work together to agree an analysis where they differed in the second pass.

7 Sampling

Lexical sense-tagging is not a well-understood task. When a task is not well-understood, it is wise to find out more about it before doing a lot of it. To find out more about it, it is necessary to look closely at it. There is too much data to look closely at everything. The approved scientific procedure, in such circumstances, is to take a sample.

We shall learn most if we use our knowledge of the domain to structure the sample. The domain can be looked at as a population of texts, or as a population of word-types, each associated with a population of tokens. For evaluating user-ready systems, the former would be appropriate, as the system would need to be able to sense-tag all words, but for an exploratory evaluation, the latter will be more informative. Our interest in a tagged corpus is for what it tells us about word-types, not for what it tells us about the texts which have been tagged. Human tagging effort will best be spent on closely investigating a sample of word-types, and, for each, examining the kind of polysemy it exhibits, and the tagging issues its corpus instances raise. We should sample the lexicon.

This section presents a sampling method in use for English SENSEVAL. The appendix includes one possible sample.

7.1 Producing a gold standard: manual tagging

In this proposal, sampling takes place at two points. First, the word types are sampled. Then, for each word type, the corpus instances are sampled. If the word *pike* is selected at the first stage, as one of the sample of types, then, for the next stage, all of its occurrences in a text corpus are identified (in the British National Corpus there are 565). A sample of, say, 200, is taken from that set. Two such instances are:[4]

> The carp and pike, which were found locally, were kitted out with lavish trimmings and served . . .
>
> Towards the close of the twelfth century the pike was used to counter cavalry charges, . . .

For pilot SENSEVAL, the sample size will be 40 word-types. A full evaluation would require several hundred. The sample size of tokens per type will range between 100 and 400, depending on the frequency and level of polysemy of the type.

For this data set to become a gold standard corpus, sense tags must be added by a person. The manual tagger's task is to say, for each of these 200 instances, whether the word is being used in its 'fish' or 'medieval weapon' sense (or neither). 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*

[4] All citations are taken from the BNC.

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 . . . [two occurrences]
>
> 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 very little previous research on this point, and none that samples systematically. The SENSEVAL pilot will shed light on the issue.

7.2 Methods: 'lexical' or 'textual'

The SEMCOR approach to tagging might be called 'textual'; human taggers work through the text, token by token. The meaning and themes of the text is foremost in the tagger's mind, and for each token to be tagged, a new set of sense-definitions is read. The approach taken in HECTOR was, by contrast, 'lexical'. The taggers worked word-type by word-type, tagging all the corpus instances for the word one after the other. In this way, the meanings and sense-distinctions of the particular word were foremost in the tagger's mind. 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). It is not possible to hold all the salient distinctions for many words in one's mind simultaneously. Taggers will make more accurate decisions faster if they work lexically rather than textually.

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. Software designed for corpus lexicography makes it easy for users to identify and account for a wide range of lexico-syntactic patterns, through flexible sorting and searching routines (Schulze and Christ, 1994).[5]

7.3 Counter-arguments

The principal counter-argument to sampling is that the resource produced will not contain any data for most words.

A further counter-argument concerns the output: with lexical sampling, it is a set of tagged contexts for each word type in the sample. Without it, it is, as in SEMCOR, a document with all words[6] tagged. This has the appeal of being simpler to conceptualise, and substantially more compact.

A further counter-argument is that word sense selections are mutually constraining, so a text like SEMCOR where the context words are disambiguated is of more value than one where they are not. Again, for evaluation purposes, this is not a concern. To the extent that a sample word can only be disambiguated correctly if other words in the sentence are also

[5] Software for the automatic discovery of such patterns, and the semi-automatic assignment of patterns to senses, is currently under development.

[6] Or, as in SEMCOR, all open-class words.

disambiguated, a WSD system which disambiguates all words will perform better than one that only attempts disambiguation of the target word.

7.4 The Sampling Scheme

The ideal sampling scheme would classify words according to the type of problems they pose to a disambiguation system, and would sample from those populations. However, the taxonomy of problem-types is not yet available (and is indeed an important intermediate goal of the exercise). The appropriate method is iterative:

- sample the lexicon according to any criteria which seem salient, and for which information is readily available
- study the sample
- feed the results of the study back into a revised sampling scheme.

The second step would involve looking only at dictionary definitions in some iterations, and looking also at corpus instances in others.

Three straightforward, available, and salient features to use for sampling are word class (eg, N, V, ADJ), frequency (as identified from a large corpus) and degree of polysemy (obtained through counting the number of senses given in a lexical resource).

Worked example

The worked example looks at nouns. Frequency and degree of polysemy were each divided into four bands, giving a sampling scheme comprising 16 classes, or cells. The resources used were WordNet version 1.5 and the BNC.

For each noun with more than fifty occurrences in the BNC, a polysemy level was established by taking the average of two figures given in WordNet, one being the number of senses in the 1978 Collins English Dictionary, the other the number of WordNet senses the word occurred in. The noun was then assigned to the appropriate cell of the sampling frame. The first number in each cell of Table 1 is the number of nouns assigned to that cell.

The frequencies for the nouns in each cell were then summed to give the second number, which is the number of word-tokens in the BNC accounted for by the types assigned to the cell (in millions).[7]

To move from these figures to a sample of nouns and of corpus instances to be tagged, two numbers are required: (1) the number of nouns from each cell are to appear in the sample, and (2) the number of corpus instances to be tagged, for each noun in the sample. The numbers of word-types selected for the sample from each cell increases with

- the number of word-types in the cell;
- the frequency-band for the cell (common words being of greater interest than rare ones).

Numbers of word-types for each cell were allocated as multiples of five, with a minimum sub-sample size of ten, with a target number of 200 for the whole sample. (For pilot SENSEVAL, there are far fewer nouns in the sample: the sample of word-types is reduced while

[7] As there are 100M words in the BNC, these are also percentages.

the number of tokens per type remains the same.) No subsamples were allocated to cells which accounted for neither many word-types nor many word-tokens.

It might seem unnecessary to take subsamples of monosemous words, since there would not appear to be a sense selection task for them. However, most words are at least occasionally used in non-standard ways, and including monosemous words would provide an opportunity for determining the scale of this phenomenon for words where the issue was not complicated by dictionary polysemy.

The number of corpus lines to be inspected per word-type in the sample, for each cell, increases with

- the frequency-band for the cell, and
- the degree of polysemy for the cell

Both these factors may be expected to give rise to a more complex pattern of word use, requiring more data to be understood. These numbers were allocated by assigning 400 corpus instances per word-type to the most frequent, most polysemous types (category AZ in the table and appendix), and reducing the figure by 40 for every step to a lower-frequency or lower-polysemy cell.

The last line of each cell of Table 1 presents, first, the proposed subsample size, for that cell, in word-types; second, the number of corpus instances to be tagged per type; and third, the product of these two numbers, representing the total number of instances to be tagged for that cell. The sum of the products across the 16 cells is 52,800, the total number of corpus instances for nouns to be tagged under this scheme.

A random sample drawn up according to this sampling scheme is presented in the Appendix.

8 Conclusion

Word sense ambiguity is a theme with implications for all corners of Language Engineering, and many people have written programs to address it. However it is not clear how good these programs are, which are better than others, which real-world problems they address, or which kinds of strategies are suitable for which kinds of cases.

One systematic way to address these questions which has borne fruit for related areas is quantitative evaluation exercise. Research groups are invited to apply their systems, all to the same data, and the performance of each system is scored by comparison with a gold standard produced by human experts. Such an exercise – pilot SENSEVAL – is curently underway, with threads for English, French and Italian.

This paper specifically addresses the preparation of the gold standard dataset against which systems are evaluated. For English, there are various manually sense-tagged datasets in existence, and they are reviewed. All have their shortcomings, but given the practical constraints of preparing an evaluation exercise on a short timescale and with limited resources, they are of great value. The HECTOR dataset is being used in English pilot SENSEVAL.

The pervasive worry in preparing the dataset is that it will not meet adequate standards of replicability: that is, if two people do the same task, they will all too frequently assign different tags to the same corpus instance. The central argument of the paper is that this is a far-reaching and difficult topic, and that a high degree of replicability can be

	Frequency band				
Num Senses	Top 200	Next 1,000	Next 5,000	Remainder	TOTALS
0–1	AW	BW	CW	DW	
Types;tokens(M)	3; .10	76; .40	1365; .94	10,141; .52	11,585; 1.96
Sample	0	10x240=2,400	15x200=3,000	20x160=3,200	45; 8,600
2–4	AX	BX	CX	DX	
Types;tokens(M)	30; .80	335; 2.27	2,436; 2.20	5,077; .40	7,878; 5.67
Sample	0	25x280=7,000	20x240=2,800	20x200=4,000	65; 3,800
5–9	AY	BY	CY	DY	
Types;tokens(M)	77; 2.20	419; 3.22	1,043; 1.22	471; .05	2,010; 6.67
Sample	10x360=3,600	25x320=8,000	15x280=420	0	50; 15,800
10+	AZ	BZ	CZ	DZ	
Types;tokens(M)	90; 3.08	170; 1.26	156; .24	29; .00	445; 4.58
Sample	15x400=6,000	15x360=5,400	10x320=3,200	0	40; 14,600
TOTALS	200; 6.18	1,000; 7.15	5,000; 4.60	15,718; .97	21,918; 18.92
	25; 9,600	75; 22,800	60; 13,200	40; 7,200	200; 52,800

Table 1: The first line in each cell names the cell. The two numbers in the second line are the number of word-types in that cell and the number of word-tokens in the BNC accounted for by those word-types in the BNC. The numbers in the third line are a proposal for how the sampling scheme should be developed. The first number is a proposal for the size of the subsample of word-types to be selected from the cell. The second is a proposal for the number of token per word-type-in-the-sample to tag. The third number, the product, is the total number of tokens to be tagged for that cell. So, for the cell AZ (representing the most common, most polysemous words); there were 90 word-types in the category, and these 90 word-types accounted for 3.08 million words in the BNC. I am proposing that 15 of these 90 words are included in the sample of word-types, and, for each of these 15 words, 400 corpus instances are tagged.

achieved, but only if the dictionary that provides the sense inventory and the individuals doing the tagging are chosen with care. The individuals best qualified to do the tagging are professional lexicographers. They have both the requisite understanding of the language and of dictionaries, and are accustomed to the discipline of working accurately and speedily through large quantities of data.

Preparing the gold standard dataset is labour-intensive, so it is important to ensure that human resources are used efficiently. To that end, sampling is a central question. The case for lexical; sampling is presented: first, words are sampled from the dictionary, then, for each sample word, the corpus instances are sampled and the tagger just tags this sample. The strongest argument for this approach concerns, again, speed and accuracy. Taggers focusing on the sense distinctions of one word, and then tagging many instances of that word, can work much faster and more accurately than ones who need to read a new dictionary entry for each word they encounter. The paper presents a detailed strategy for making optimum use of tagger effort.

In pilot SENSEVAL, these ideas are being explored as far as time and resources allow. We are closely examining progress in the pilot, and look forward to applying more sophisticated versions of the model, on a larger scale, in the future.

Appendix: A sample of English nouns

For decoding category names see Table 1. Words for which the BNC did not provide as many corpus instances as indicated in the table have been excluded.

AY woman support city bank government member effect father moment relationship

AZ year back difference light community law way man mother court use sense group authority face

BW mouth american restaurant chest discussion employee spokesman manufacturer leadership awareness

BX opportunity context adult noise ball conservative cash while proportion bill membership gift expense republic penalty drink employment ratio knife championship category son shop corporation efficiency

BY cloud officer energy arrangement winter cheek engineer daughter code institution recovery minority works competition region introduction magazine examination phase chip ring bread move village mechanism

BZ pattern pair impression hole supply height flight truth key reader preparation standard heart representation metal

CW spreadsheet european dumping zoo snag chap consonant adequacy londoner rejection broadcaster hospice colliery layout plight

CX armament dice keeping contractor statistics deletion hurry referee porch loom leisure semantics prohibition granite thickness motherhood essential magnate innovation melon

CY tariff priest dive reservoir favour trumpet cry mortar slate fraction synthesis pet curfew distortion mail

CZ fellow spread knot discharge bolt puff jump grip float stroke

DW sac kerb qualifying humanist animosity airframe mystic chum anemone dick rectum tenet marshall raisin priory prairie eec blazer operand smog

DX sister-in-law upturn deformation absentee chub buttock mousse kinsman sunrise vestige glint rye feud mercenary pauper tycoon miniature devotee junta backlog

References

asnd Lin Chase, Steven J. Young. 1998. Speech Recognition Evaluation: A Review of the US CSR and LVCSR Programmes. *Computer Speech and Language*, this issue.

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

Bruce, Rebecca and Janyce Wiebe. 1994. Word sense disambiguation using decomposable models. In *Proc. 32nd Annual Meeting of the ACL*, pages 139–145, Las Cruces, New Mexico. ACL.

Bruce, Rebecca and Janyce Wiebe. 1998. Word sense distinguishability and inter-coder agreement. In *Proc. 3rd Empirical Methods in Natural Language Processing*, pages 53–60, Granada. ACL SIGDAT.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Choueka, Yaacov and Serge Lusignan. 1985. Disambiguation by short contexts. *Computers and the Humanities*, 19:147–157.

Church, Kenneth. 1992. Current practice in part of speech tagging and suggestions for the future. In Simmons, editor, *Sbornik praci: In Honour of Henry Kučera*. Michigan Slavic studies.

CIDE, 1995. *Cambridge International Dictionary of English*. CUP, Cambridge, England.

Clear, Jeremy. 1994. I can't see the sense in a large corpus. In Ferenc Kiefer, Gabor Kiss, and Julia Pajzs, editors, *Papers in Computational Lexicography: COMPLEX '94*, pages 33–48, Budapest.

Cowie, Jim, Joe Guthrie, and Louise Guthrie. 1992. Lexical disambiguation using simulated annealing. In *COLING 92*, pages 359–365, Nantes.

Dagan, Ido and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

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

Fellbaum, Christiane, Joachim Grabowski, Shari Landes, and Andrea Baumann. 1996. Matching words to senses in WordNet: Naive *vs.* expert differentiation of senses. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press, Cambridge, Mass.

Gale, William, 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.

Harley, Andrew and Dominic Glennon. 1997. Combining different tests with additive weighting and their evaluation. In Marc Light, editor, *Tagging Text with Lexical Semantics: Why, What and How?*, pages 74–78, Washington, April. SIGLEX (Lexicon Special Interest Group) of the ACL.

Hirschman, Lynette. 1998. The Evolution of Evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language*, this issue.

Hirst, Graeme. 1987. *Semantic Interpretation and the Resolution of Ambiguity*. CUP, Cambridge, England.

Ide, Nancy and Jean Véronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

Jorgensen, Julia C. 1990. The psychological reality of word senses. *Journal of Psycholinguistic Research*, 19(3):167–190.

Karov, Yael and Shimon Edelman. 1998. Similarity-based word sense disambiguation. *Computational Linguistics*, 24(1):41–60.

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

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

Kilgarriff, Adam. 1997. What is word sense disambiguation good for? In *Proc. Natural Language Processing in the Pacific Rim (NLPRS '97)*, pages 209–214, Phuket, Thailand, December.

Krippendorf, Klaus. 1980. *Content Analysis: an introduction to its methodology*. Sage Publications.

Lesk, Michael E. 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.

Light, Marc, editor. 1997. *Tagging Text with Lexical Semantics: Why, What and How?*, Washington, April. SIGLEX (Lexicon Special Interest Group) of the ACL.

Lin, Dekang. 1997. Usin syntactic dependency as local context to resolve word sense ambiguity. In *Proc. 35th Annual Meeting of the ACL and 8th Conference of the EACL*, pages 64–71, Madrid, July. ACL.

Miller, George. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography (special issue)*, 3(4):235–312.

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

Resnik, Philip 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.

Rigau, German, Jordi Atserias, and Eneko Agirre. 1997. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proc. 35th Annual Meeting of the ACL and 8th Conference of the EACL*, pages 48–55, Madrid, July. ACL. CMP-LG 9704007.

Samuelsson, Christer and Atro Voutilainen. 1997. Comparing a linguistic and a stochastic tagger. In *Proc. 35th Annual Meeting of the ACL and 8th Conference of the EACL*, pages 246–253, Madrid, July. ACL.

Schulze, Bruno and Oliver Christ, 1994. *The IMS Corpus Workbench*. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Schütze, Hinrich. 1998. Word sense discrimination. *Computational Linguistics*, 24(1):97–124.

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

Stock, Penelope F. 1983. Polysemy. In *Proc. Exeter Lexicography Conference*, pages 131–140.

Vossen, Piek, Geert Adriaens, Nicoletta Calzolari, Antonio Sanfilippo, and Yorick Wilks, editors. 1997. *Automatic Information Extraction and Building Lexical Resources for NLP applications*, Madrid, July. ACL and EC Projects EuroWordnet, Sparkle and Ecran.

Wiebe, Janyce, Julie Maples, Lei Duran, and Rebecca Bruce. 1997. Experience in WordNet sense taggign in the Wall Street Journal. In Marc Light, editor, *Tagging Text with Lexical Semantics: Why, What and How?*, pages 8–11, Washington, April. SIGLEX (Lexicon Special Interest Group) of the ACL.

Wilks, Yorick, Brian M. Slator, and Louise Guthrie. 1996. *Electric words: dictionaries, computers and meanings*. MIT Press, Cambridge, Mass.

Wilks, Yorick and Mark Stevenson. 1997. Sense tagging: semantic tagging with a lexicon. In Marc Light, editor, *Tagging Text with Lexical Semantics: Why, What and How?*, pages 47–51, Washington, April. SIGLEX (Lexicon Special Interest Group) of the ACL.

Yarowsky, David. 1992. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *COLING 92*, Nantes.

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