Computers and the Humanities **34:** 1–13, 2000.

Introduction to the Special Issue on SENSEVAL

A. KILGARRIFF[1] and M. PALMER[2]
[1]*ITRI, University of Brighton;* [2]*University of Pennsylvania*

Abstract. SENSEVAL was the first open, community-based evaluation exercise for Word Sense Disambiguation programs. It took place in the summer of 1998, with tasks for English, French and Italian. There were participating systems from 23 research groups. This special issue is an account of the exercise. In addition to describing the contents of the volume, this introduction considers how the exercise has shed light on some general questions about word senses and evaluation.

Key words: word sense disambiguation, evaluation, SENSEVAL

1. Introduction

SENSEVAL was the first open, community-based evaluation exercise for Word Sense Disambiguation programs. It took place in the summer of 1998 under the auspices of ACL SIGLEX (the Association for Computational Linguistics Special Interest Group on the Lexicon), EURALEX (European Association for Lexicography), ELSNET, and EU Projects SPARKLE and ECRAN. This special issue is an account of the exercise.

In this introduction, we first describe the problem and the historical context; then the papers; then we address some criticisms of the evaluation paradigm; and finally, we look forward to future SENSEVALs.

2. SENSEVAL: The Context

2.1. THE PROBLEM

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 an issue for people. We are very rarely slowed down in our comprehension by the need to determine which meaning of a word applies. But it is a very difficult task for computers. The clearest case is in Machine Translation. If the English word *drug* translates into French as either *drogue* ('bad' drugs) or *médicament* ('good' drugs), then an English-French MT system needs to disambiguate *drug* if it is to make the correct translation. Similarly, information retrieval systems may retrieve documents about a *drogue* when the item of interest is a *médicament*; information extraction systems may make wrong assertions; text-

to-speech systems will make errors where there are multiple pronunciations for the same spelling, as in violin *bows* and ships' *bows*. For virtually all Natural Language Processing applications, word sense ambiguity is a potential source of error.

For forty years now, people have been writing computer programs to do Word Sense Disambiguation (WSD). The field is surveyed, from earliest times to recent work, in (Ide and Véronis, 1998) and the reader is directed to that paper for historical background and the kinds of methods that have been used.

2.2. WHAT ARE WORD SENSES?

Before a WSD problem is well-defined, a set of word senses to disambiguate between is required. This raises a number of issues. First, which dictionary? People often refer to 'the dictionary' as if there were just one, definitive one. But dictionaries differ and, for very many words, any two will give different analyses. Readings treated as distinct in one dictionary will be merged in the other. Bigger dictionaries will give more senses than smaller ones. Lexicographic policies regarding grammar, phraseology and metaphor all affect what a particular dictionary treats as a sense or subsense. Also, some dictionary entries are better than others. Sometimes the lexicographer will not have arrived at a clear image of what the distinction between two putative senses is before writing the entry, and sometimes, even though the distinction was clear to him/her, he or she will not have succeeded in making in clear in the entry.

Second, homonymy and polysemy. In homonymy, there are two or more distinct 'words' which happen to have the same form. In polysemy a single word has multiple meanings. Distinctions between homonyms are clear, and disambiguating between them is, for people, straightforward. For polysemous words, it may not be so, either in the abstract or in relation to particular contexts. When a drug is stolen from the pharmacy, it is indeterminate between *drogue* and *médicament*. It might appear appealing to distinguish homonymy resolution from polysemy resolution, but in practice, there are no general, systematic methods for making the distinction, and experts frequently disagree.

While relations between homonyms are arbitrary, relations between polysemes are riddled with regularities. Thus *rabbit* is like *chicken*, *turkey* and *lamb* in having both an 'animal' sense and a 'meat of that animal' sense. *Kangaroo* and *emu* also appear to participate in the pattern; certainly, one might find either on a restaurant menu with a 'meat' reading required. Where a regularity could be applied to a word, but the derived sense is neither particularly common, nor is there anything about it which is not predictable, it will not generally be listed in a dictionary and we may say it is not 'lexicalised'. Yet clearly, words are used in such ways and a disambiguation program will need to do something with them.

Also, the regularities are rarely fully predictive. *Pig* does not have the meat sense.

In sum, there are various reasons why people who do not have any trouble understanding a word in context, might nonetheless have difficulty assigning it to a sense from a dictionary. In some cases, towards the homonymy end of the spectrum, the word sense disambiguation problem does appear to map straightforwardly to something that people do when they understand a sentence with an ambiguous word in it. As we move towards senses that are closely related, the task seems more artificial, and people may disagree. We return to the causes and implications of such disagreements at various points in this introduction and elsewhere in the special issue.

2.3. EVALUATION

There are now many 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 DARPA competitive evaluations for speech recognition, dialogue systems, information retrieval and information extraction has been that the focus provided by an evaluation brings 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 (see, e.g. Gaizauskas, 1998).

Reaping these benefits involves overcoming two major hurdles. The first is agreeing an explicit and detailed definition of the task. The second is producing 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 program is to disambiguate, the "sense inventory" problem. Producing a gold standard corpus for WSD is both expensive, as it requires many person-months of annotator effort, and hard because, as earlier evidence has shown, if the exercise is not set up with due care, different individuals will often assign different senses to the same word-in-context.

2.4. HISTORY OF WSD EVALUATION

People producing WSD systems have always needed to evaluate them. A system developer needs a test set of some sort to determine when the system is working at all, and whether a change has improved matters or made them worse. So system developers have frequently worked through a number of sentences containing the words of interest, assigning to each a sense-tag from whatever dictionary they were using. They have then, on some occasions, stated the percentage correct for their system in the write-up.

Gale, Church and Yarowsky (1992) review, exhaustively and somewhat bleakly, the state of affairs as at 1992. They open with:

> We have recently reported on two new word-sense disambiguation systems . . . [and] have convinced ourselves that the performance is remarkably good. Nevertheless, we would really like to be able to make a stronger statement, and therefore, we decided to try to develop some more objective evaluation measures.

First they compare one of their systems' (Yarowsky, 1992) performance with that of other WSD systems for which accuracy figures are available (taking each word addressed by each other system in turn). While the comparison of numbers suggests in most cases that their system does better, they note

> one feels uncomfortable about comparing results across experiments, since there are many potentially important differences including different corpora, different words, different judges, differences in precision and recall, and differences in the use of tools such as parsers and part of speech taggers etc. In short, there seem to be a number of serious questions regarding the commonly used technique of reporting percent correct on a few words chosen by hand. Apparently, the literature on evaluation of word-sense disambiguation fails to offer a clear model that we might follow in order to quantify the performance of our disambiguation algorithms. (p. 252)

The paper was written at a time of increasing interest in evaluation in Language Engineering in general, and the concerns they list are in large part those that are resolved by collaborative, co-ordinated community-wide evaluation exercises as in the DARPA model.

The topic was raised again four years later, as the central issue of a workshop of the ACL Lexicon Special Interest Group (SIGLEX) in Washington, April 1997. The DARPA community had been baffled by the difficulty, perhaps impossibility, of determining a methodology for the evaluation of semantic interpretation. There was not even a consensus on the right level of semantic representation, let alone what that representation should contain. Martha Palmer, as chair of SIGLEX, suggested that a workshop be organised around the central questions of whether or not "hand tagged text [would] also be of use for assigning semantic characteristics to words in their context . . . to what end should hand tagging be performed, what lexical semantic information should be hand tagged, and how should this tagging be done?" During the workshop, chaired by Marc Light, sense tagging was recognised as a relatively uncontroversial level of semantic analysis that might be more amenable to evaluation than other more problematic levels. Resnik and Yarowsky made some practical proposals for evaluation of WSD systems using machine learning techniques (Resnik and Yarowsky, 1997). These were broadly welcomed, and led to extensive and enthusiastic discussions. There was a high degree of consensus that the field of WSD would benefit from careful evaluation,

and that researchers needed to collaborate and make compromises so that an evaluation framework could be agreed upon. An actual experiment in a community wide-evaluation exercise would allow us to address three fundamental questions:

1. What evidence is there for the 'reality' of sense distinctions?
2. Can we provide a consistent sense tagged Gold Standard and appropriately measure system performance against it?
3. Is sense tagging a useful level of semantic representation: what are the prospects for WSD improving overall system performance for various NLP applications?

Following the Washington meeting, Adam Kilgarriff undertook the co-ordination of a first evaluation exercise, christened SENSEVAL.[1] The exercise culminated in a workshop (held at Herstmonceux Castle, Sussex, England) in September 1998. Most of the papers in this special issue have their origins in presentations at that workshop. The evidence of the workshop sheds light on the first question, and gives an unequivocal 'yes' to the second. The third is more complex, and we return to it in Section 4.

3. Papers

3.1. LANGUAGES COVERED; 'FRAMEWORK' PAPERS

Most research in WSD has been on English. There are many resources available for English, much commercial interest, and much expertise in the problems it presents. It is easiest to set up an exercise for English. However, there was no desire for hegemony, so ACL SIGLEX's position was simply that, wherever there was an individual or group with the commitment and resources to set up an exercise for a given language, they would be welcomed and encouraged, though they would then be responsible for all the language-specific work (including funding the resource development). There were preliminary discussions regarding six languages in all, and for the first SENSEVAL, there were English, French and Italian tasks. The French and Italian teams worked together under the banner of ROMANSEVAL and adopted parallel designs. For each of the three exercises, there is a paper describing how the exercise was set up, and the results: for English, by Kilgarriff and Rosenzweig; for French, by Segond; and for Italian, by Calzolari and Corazzari. These papers describe the choice of lexicon and corpus for each task; the methods used for choosing a sample of word types; the approach to manual sense tagging; the level of agreement between different human sense-taggers; baselines; system results; and problems and anomalies encountered during the whole process.

An evaluation needs a scoring metric, and one of the issues raised by (Resnik and Yarowsky 1997) was that a simple metric, whereby a correct response scores 1 and anything else scores 0, is not satisfactory. It says nothing about what to do where there are multiple correct answers, or where a system returns multiple responses, or where the tags are hierarchically organised, so that one tag may be a generalisation or specialisation of another. In the one paper in the special

Table I. Numbers of participants for each language

	Systems	Research groups	Papers	Brief note
English	18	17	15	3
French	5	4	1	3
Italian	2	2	1	0
Totals	25	23	17	6

issue which is not specific to WSD, Melamed and Resnik present a scoring scheme meeting the desiderata. The scheme underlay the scoring strategies used in SENSEVAL.

Krishnamurthy and Nicholls describe the process of manually tagging the English test corpus, with detailed discussion of the cases where the lexical entry and/or corpus instance meant that there was not a straightforward, single, correct sense tag for the corpus instance. They thereby provide a research agenda for work in the area: what must one do, to the dictionary, or WSD system, or larger theoretical framework, to not inevitably go wrong, for each of these types of cases?

In a short note, Moon asks what the scale of the WSD problem is, and shows that it relates, for general English, to the order of 10,000 words – a consideration that becomes critical should it be necessary to do lexicographical work on each one of those words.

3.2. PARTICIPATING SYSTEMS

All research teams which participated in the evaluation – that is, which applied their WSD system to the test data and returned results – were invited to submit descriptions of their system and its performance on the task to the special issue. Table I shows, for each task, how many participating systems, research groups and special issue papers there are.[2]

For most of the 25 participating systems, there is a paper in the special issue (and for six of the remainder, there are brief descriptions inserted as appendices to the appropriate 'framework' paper).

The systems use a range of machine learning algorithms and consult a variety of lexical resources. When this exercise was first proposed, in Washington in 1997, it was notable that the participants fell into opposing camps – the proponents of machine learning techniques versus the proponents of hand-crafted lexical resources. Each camp eagerly anticipated demonstrating their superiority in SENSEVAL. Notable at the workshop was the frequency with which participants had merged the two approaches. Several 'unsupervised systems' – those relying on lexical resources – made extensive use of the training data to fine-tune their

systems, and several 'supervised systems' – those relying on machine learning from training data – had a lexical resource as a fall-back where the data was insufficient. When it came to getting the task done, the purity of the approach was less important than the robustness of the system performance. The extensive discussion of criteria for a sense inventory also created more awareness among the participants of how fundamental the lexicon is to the task. It is only worth learning sense distinctions if they can in fact be distinguished.

The English exercise was set up with substantial amounts of training data, which supported machine-learning approaches. This was clearly reflected in the results, with the machine learning systems performing best. The highest performing systems utilised a wide range of features, including inflectional form of the word to be disambiguated, part-of-speech tag sequences, semantic classes, and collocates at specific positions as well as 'anywhere in a *k*-word window of the target word'. Some of these features are dependent on others, so techniques such as O'Hara et al.'s that do not assume independence when incorporating features, could make a more principled use of the data. This makes the good performance of Chodorow et al. intriguing as their Bayesian model does assume independence. One system (Hawkins's) used some manually rather than automatically derived features, with the manual acquisition organised so that it could be rapidly bootstrapped from untagged training material.

Veenstra et al. improved their system performance when they optimised the settings in their model for each individual word based on performance in a cross validation exercise. They got quite distinct settings for each individual lexical item. Approaches that are sensitive to such individual differences are clearly necessary, but the requisite amount of training data is disconcerting. An ability to leverage sparse data effectively, as was done by exemplar based approaches, mitigates this need to some degree.

One of the pleasant outcomes of the evaluation was that many groups were clearly using the data to test a particular attribute of their system, rather than focusing simply on maximising results. Systems that used only grammatical relations or subcategorisation frames did not fare as well in the performance comparisons, but gained valuable information about the contribution of individual feature types. This type of scholarly approach to training and testing benefits the field as much as an approach that is primarily focused on winning the bake-off. Future SENSEVALS will do well to continue to foster this exploratory attitude.

3.3. DISCUSSION PAPERS

The papers by Hanks, Palmer, Ide, and Wilks examined the fundamental question of how sense distinctions can be made reliably, providing critical perspectives and suggestions for future tasks. The question of the role of WSD in a complete NLP system is also raised.

Hanks asks, simply, "Do word meanings exist?" and reminds us of the extent to which they are figments of the lexicographer's working practice. As he says, "if senses don't exist, then there is not much point in trying to disambiguate them". His corpus analyses of *bank*, *climb* and *check* show how different components of the meaning potential of the word are activated in different contexts. His paper is a call for representations of word meaning that go beyond "checklist theories of meaning" and record meaning components, organised into hierarchies and constellations of prototypes, and for algorithms that work out which of the components are activated in a context of use.

The Palmer paper is complementary, in that it asks the same question but from the perspective of an NLP system. How are different senses of the same word characterised in a computational lexicon? She focuses on verb entries. Since they typically consist of predicate argument structures with possible semantic class constraints on the arguments, possible syntactic realizations and possible inferences to be drawn, alternative senses must differ concretely in one or more of these aspects. The more closely each entry in a dictionary "checklist" can be associated with a concrete change along one or more of these dimensions, the more readily a computational lexicon can capture the relevant distinctions. The meaning components desired by Hanks can correspond to one or more elements of this type of representation, suggesting a measure of convergence between the lexicographic community and the computational lexical semantics community.

Ide presents a study into the use of aligned, parallel corpora for identifying word senses as items that get systematically translated into one or more other languages in the same way. This is a highly appealing notion, and is indeed a strategy used by lexicographers in determining the senses a word has in the first place. It offers the prospect of taking the confounding factors of lexicographic practice out of the definition of word senses. Ide's study is small-scale, but charts the issues that would need addressing if the strategy was to be adopted more widely (see also section 5 below).

Wilks asks several central questions about the way in which the WSD field is proceeding: will data-driven methods reach their upper bound all too soon, precipitating a return to favour of AI strategies? Where do discussions of lexical fields and vagueness take us? He presents the case against the "lexical sample" aspect of the design of the SENSEVAL task.[3] He also addresses the larger question of the usefulness of WSD for complete NLP systems and notes that Kilgarriff is associated with a sceptical view, which sits oddly for one organising SENSEVAL:

> There need be no contradiction there, but a fascinating question about motive lingers in the air. Has he set all this up so that WSD can destroy itself when rigorously tested? . . . [the issue goes] to the heart of what the SENSEVAL workshop is for: is it to show how to do better at WSD, or is it to say something about word sense itself?

Let me (Kilgarriff) take this opportunity to respond. SENSEVAL is, from one point of view, an experiment designed to replace scepticism about both the reality

of word senses and the effectiveness of WSD, by percentages. It answers some simple, quantitative questions: what is the upper bound for human inter-tagger-agreement (95%); and at what level do state-of-the-art systems perform (75–80%) (both answers relative to a fine-grained, corpus-based dictionary; see Kilgarriff and Rosenzweig, this volume, for discussion). SENSEVAL provided a clear picture of the types of systems that performed best (the 'empiricist' methods, using as much training data as was available) and, as a side-product, provided an extensive sense-tagged corpus where instances that had given rise to tagger disagreement could be identified for further analysis (Kilgarriff, 2000).

We return to the relation between SENSEVAL and the usefulness of WSD in complete NLP systems in the next section.

4. Responses to Criticisms

Given our conscious similarity to the DARPA quantitative evaluation paradigm, the recurring criticisms of it are the first ones to be addressed. These are as follows:[4]

1. It discourages novel approaches and risk taking, since the focus is on improving the error rate. This can be done most reliably by duplicating the familiar methods that are currently scoring best.
2. There is a substantial overhead involved both in setting up the evaluations and in participating in them.
3. It encourages a competitive (as opposed to collaborative) ethos.
4. Unless the tasks are carefully chosen to focus on the fundamental problems in the field, they will draw energy away from those problems.

The first criticism cannot hold of a first evaluation of a given task (and is unlikely to apply unless the evaluation becomes a substantial undertaking with reputations hanging on the outcome). Indeed, the informal flavour of SENSEVAL fostered experimentation and diversity. The second also does not apply to this first, small-scale evaluation (where much was done on goodwill) but is likely to apply to future, hopefully larger-scale evaluations. The case will have to be made for the substantial costs reaping commensurable benefits. There are of course many precedents for this; as (Hirschman, 1998) says,

> Evaluation is itself a first-class research activity: creation of effective evaluation methods drives rapid progress and better communication within a research community. (pp. 302–303)

The third is a concern that was discussed at length in the course of SENSEVAL, particularly in relation to the question, should the full set of results be made public? This would potentially embarrass research teams whose systems did not score so well, and may deter people from participating in the future. It was eventually agreed that, given the early stage of maturity of the field, the merits of having all results in the open outweighed the risks, but not without dissenters. In more general terms, our experience has been that of other DARPA evaluations: both the fellow-feeling

that comes of working on the same problem and the modest dose of competitive tension have been productive.

The last criticism demands much fuller discussion, and lies at the heart of evaluation design. It was the third fundamental question that we were hoping to address: *Is sense tagging a useful level of semantic representation: what are the prospects for WSD improving overall system performance for various NLP applications?*

One critic of the process chose not to participate because, in their system, WSD occurred as a byproduct of deeper reasoning. It would not make sense to participate in an exercise that treated WSD as of interest in its own right. They were engaged in a harder task, so had no inclination to work on intermediate outputs as defined by an easier task. The sense distinctions that needed making would also only be identified in the course of specifying the overall NLP system outputs, so, taking them from a dictionary was not a relevant option (see also Kilgarriff, 1997).

The question recurs in the evaluation literature, as, for any subtask, the validity of evaluation is contingent on the validity of the analysis that identifies the subtask as a distinct process (Palmer et al., 1990; Sparck Jones and Galliers, 1996; Gaizauskas, 1998). Despite being theory-dependent in this way, subtask evaluations can clearly be of great value. Evaluations focused on end results (which are often also user-oriented) tend not to help developers determine the contributions of individual components of a complex system. Thus parsing is generally agreed upon as a separable NLP task, and evaluations associated with the Penn Treebank have emphasised syntactic parsing as a separate component. The focus has resulted in significantly improved parsing performance (even though re-integrating these improved parsers into NLP applications is itself a non-trivial task that has yet to be achieved).

SENSEVAL can be seen as an experiment to test the hypothesis that "WSD is a separable NLP subtask". It would seem some parts of the task, such as homograph resolution, can be effectively addressed with nothing more than shallow-processing WSD techniques, while others, such as metaphor resolution, require full-fledged NLP. Results suggest that at least 75% of the task could usefully be allocated to a shallow-processing WSD module, and that at least 5% could not.

Although we may have demonstrated that WSD can be defined as a separate task, we have not established that good WSD performed as a separate stage of processing can improve the overall performance of an NLP application such as IR or MT. Indeed, the difficulty of demonstrating the positive impact of natural language processing subcomponents on Information Retrieval has dogged the field for decades. These subcomponents, whether they perform noun phrase chunking or WSD, may show improved performance on their individual subtasks, but they have little effect on the overall task performance (Buckley and Cardie, 1998; Voorhees, 1999). Machine Translation and cross-linguistic IR would seem more promising areas for illustrating the benefit of WSD. A clear demonstration would require establishing the baseline performance of a given NLP system, and then showing a significant percentage improvement in those figures when WSD is added. For

instance, specific lexical items can be highlighted in a Machine Translation task, and the number of errors in translation of these items both with and without WSD calculated. Future SENSEVALs must address this issue more directly.

5. Towards Future SENSEVALs

SENSEVAL participants were enthusiastic about future SENSEVALs, with several provisos. Some wanted evaluation on texts with all content words tagged. General NLP systems that perform WSD on the route to a comprehensive semantic representation need to disambiguate every word in the sentence, so, for people with this goal on their medium-term horizon, an evaluation which looked only at corpus instances of selected words missed the central issue. Also, it seems likely that tag-assignments are mutually constraining. Only data with tags for several of the words in each sentence can pinpoint the interactions. A pilot study for the tagging of running text with revised WordNet senses was presented at SIGLEX99 and positively received (Palmer et al., 2000).

Participants also wanted confirmation that the senses they were distinguishing were relevant to some type of NLP task, such as Information Retrieval or Machine Translation. (There is a close overlap between this concern and the goal of confirming WSD as a separable NLP subtask, as discussed above.) At the Herstmonceux workshop, we resolved to tie WSD more closely to Machine Translation, and to attempt to use sense inventories which were appropriate for Machine Translation tasks. The foundational work of Resnik and Yarowsky (1997, 1999) and Ide (this volume) on clustering together monolingual usages based on similar translations provides a preliminary framework. It is of course well known that languages often share several senses for single lexical items that are translations of each other, and translation simply preserves the ambiguities. Conversely, different translations in another language do not always correlate with a valid sense distinction in the source language (Palmer and Wu, 1995). Having the same translation does not ensure sense identification, and having separate translations does not ensure sense distinctions. However, multiple translations of a single word can provide objective evidence for possible sense distinctions, and, given our current state of knowledge, any such evidence is to be embraced.

6. Conclusion

This special issue provides an account of SENSEVAL, the first open, community-based evaluation for WSD programs. There were tasks for three languages, and 23 research teams participated. By making direct comparisons between systems possible, and by forcing a level of agreement on how the task should be defined, the exercise sharpened the focus of WSD research.

The volume contains detailed accounts of how the evaluation exercises were set up, and the results. Most of the participating systems are described and there are

position papers on several of the difficult issues surrounding WSD and its evaluation: what word senses are, how they should be identified, and how separable from a particular application context the WSD task, and any specific sense inventory, will ever be. As this introduction conjectures, for some of these questions, the outcomes from SENSEVAL can be seen as quantitative answers.

We hope that SENSEVAL, and this volume, will provide a useful reference point for future SENSEVALs and other future WSD research worldwide.

Acknowledgements

We would like to thank Cambridge University Press, EPSRC (grant M03481), ELRA (European Linguistic Resources Association), the European Union (DG XIII), Longman Dictionaries and Oxford University Press for their assistance in goods and kind with the SENSEVAL exercise. We would also like to thank Carole Tiberius for her role in organising the workshop.

RESOURCES AVAILABLE, SEE WEBSITE

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

Notes

[1] The name is due to David Yarowsky.

[2] For the purposes of this table, 'research teams' are treated as distinct if they are responsible for different systems, and the different systems have different writeups, even if the individuals overlap.

[3] For the case for the lexical sample approach, see section 2 of Kilgarriff and Rosenzweig, this volume.

[4] For discussion see Sproat et al. (1999).

References

Buckley, C. and C. Cardie. "EMPIRE and SMART Working Together". Presentation at the DARPA/Tipster 24-Month Meeting, 1998.

Gaizauskas, R. "Evaluation in Language and Speech Technology: Introduction to the Special Issue". *Computer Speech and Language*, 12(4) (1998), 249–262.

Gale, W., K. Church and D. Yarowsky. "Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs". In *Proceedings, 30th ACL*. 1992, pp. 249–256.

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

Ide, N. and J. Véronis. "Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art'. *Computational Linguistics*, 24(1) (1998), 1–40.

Kilgarriff, A. "Foreground and Background Lexicons and Word Sense Disambiguation for Information Extraction". In *Proc. Workshop on Lexicon Driven Information Extraction*. Frascati, Italy, 1997, pp. 51–62.

Kilgarriff, A. "Generative Lexicon Meets Corpus Data: The Case of Non-Standard Word Uses". In *Word Meaning and Creativity*. Ed. P. Bouillon and F. Busa, Cambridge: Cambridge University Press, forthcoming, 2000.

Palmer, M., H.T. Dang and J. Rosenzweig. “Sense Tagging the Penn Treebank”. Submitted to the *Second Language Resources and Evaluation Conference*. Athens, Greece, 2000.

Palmer, M., T. Finin and S. Walters. “Evaluation of Natural Language Processing Systems”. *Computational Linguistics*, 16(3) (1990), 175–181.

Palmer, M. and Z. Wu. “Verb Semantics for English-Chinese Translation”. *Machine Translation*, 10, (1995), 59–92.

Resnik, P. and D. Yarowsky. “A Perspective on Word Sense Disambiguation Methods and Their Evaluation”. In *Tagging Text with Lexical Semantics: Why, What and How?* Ed. M. Light, Washington, 1997, pp. 79–86.

Resnik, P. and D. Yarowsky. “Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation”. *Natural Language Engineering Journal*, to appear.

Sparck Jones, K. and J. Galliers. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Berlin: Springer-Verlag, 1996.

Sproat, R., M. Ostendorf, and A. Hunt: 1999, “The Need for Increased Speech Synthesis Research”. Report of the 1998 NSF Workshop for Discussing Research Priorities and Evaluation Strategies in Speech Synthesis.

Voorhees, E.M.: 1999, “Natural Language Processing and Information Retrieval”. In *Proceedings of Second Summer School on Information Extraction*. Springer-Verlag, Lecture Notes in Artificial Intelligence.

Yarowsky, D. “Word-Sense Disambiguation Using Statistical Models of Roget’s Categories Trained on Large Corpora”. In *COLING 92*. Nantes, 1992.