Natural Language Engineering **8** (4): 279–291. © 2002 Cambridge University Press
DOI: 10.1017/S1351324902002966 Printed in the United Kingdom

Introduction to the special issue on evaluating word sense disambiguation systems

PHILIP EDMONDS
Sharp Laboratories of Europe, Oxford Science Park, Oxford OX4 4GB, UK
e-mail: phil@sharp.co.uk

ADAM KILGARRIFF
Information Technology Research Institute, University of Brighton,
Lewes Road, Brighton BN2 4GJ, UK
e-mail: Adam.Kilgarriff@itri.brighton.ac.uk

(*Received 5 August 2002*)

1 Assessing performance on word sense disambiguation

Has system performance on Word Sense Disambiguation (WSD) reached a limit? Automatic systems don't perform nearly as well as humans on the task, and from the results of the SENSEVAL exercises, recent improvements in system performance appear negligible or even negative. Still, systems do perform much better than the baselines, so something is being done right. System evaluation is crucial to explain these results and to show the way forward. Indeed, the success of any project in WSD is tied to the evaluation methodology used, and especially to the formalization of the task that the systems perform. The evaluation of WSD has turned out to be as difficult as designing the systems in the first place.

This special issue of *Natural Language Engineering* explores the evaluation of WSD systems with particular reference to SENSEVAL. SENSEVAL-1, in 1998, was the first open, community-based evaluation exercise for WSD programs, and SENSEVAL-2, in 2001, was the second. Both were organized by ACL-SIGLEX, the Association for Computational Linguistics Special Interest Group on the Lexicon. We first describe the problem, the critical issues, and the history. We then introduce the papers. To conclude, we look forward to future SENSEVALS.

2 WSD and its evaluation

2.1 The problem

Most common words have more than one meaning, but when a word is used, just one of those meanings will apply, generally speaking. People are very rarely slowed down in their comprehension by the need to consciously determine the meaning that applies. However, it is very difficult to formalize this process of disambiguation, which is required in many applications of language technology. Take Machine Translation (MT). If the English word *drug* translates into French as either *drogue* or *médicament*, then an English-French MT system needs to disambiguate *drug* to make the correct translation. Similarly, information retrieval systems may erroneously retrieve documents about an illegal narcotic when the item of interest is a medication; analogously, information extraction systems may make wrong assertions; text-to-speech systems will confuse violin bows and ships' bows. For virtually all applications of language technology, word sense ambiguity is a potential source of error.

For 40 years now, people have been writing computer systems to do WSD. The field is broadly surveyed by Ide and Véronis (1998), and several recent textbooks (Manning and Schütze (1999) and Jurafsky and Martin (2000)) provide more historical background and describe the kinds of algorithms that have been used.

2.2 Evaluation

US DARPA common evaluations for applications of language technology such as speech-to-text, dialogue systems, information retrieval (TREC), information extraction (MUC), and text summarization (DUC) have been very successful in stimulating rapid scientific progress. They have brought the research community to consensus on appropriate tasks for evaluation, have designed metrics for measuring comparative performance and for diagnosing system strengths and weaknesses, and have led to the development of common, open, resources.

To reap these benefits for WSD, the research community must overcome two major hurdles. The first is to agree an explicit and detailed definition of the task. Defining the task includes identifying for each word a set of senses between which a program is to disambiguate: the 'sense inventory' problem. The second hurdle is to produce a 'gold standard' corpus of correct answers. For WSD, this 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.

2.3 The sense inventory

A sense inventory partitions the meaning of each word into its senses. Only a good sense inventory can be a valid approximation to the truth about the lexicon, but perhaps even the best possible inventory is woefully inadequate. This is because what counts as a sense is notoriously difficult to define. Different applications require different sorts of distinctions. For example, an ambiguity that is preserved in translation (e.g. *interest* from English to French) does not need to be broken down, whereas in information extraction it would have to be. This is reflected in the differences between monolingual and bilingual dictionaries, and thesauri, which split senses along different lines. And then, Pustejovsky (1995) argues that senses cannot even be enumerated, working from the generative lexicon paradigm. Or maybe the very idea of word sense is suspect, with corpus data frequently revealing loose and overlapping categories of meaning, and standard meanings for words extended and exploited in a bewildering variety of ways (Kilgarriff 1997; Hanks 2000). Theoretical positions may point out how such problems will sometimes arise, but it takes a more extensive study to quantify whether they are actual obstacles for language engineering (Kilgarriff 2001).

Many questions arise in choosing a sense inventory. Are the senses well-motivated and attested in the corpus? Are the senses too fine-grained (too much splitting) or two coarse-grained (too much lumping)? Does the inventory reflect the right domain and genre of language to be tested? How are senses described, differentiated, and organized in the resource? Can lexicographers or others tag with the required consistency and replicability? An evaluation task requires a strategy on each of these questions, and one is expounded in detail in Kilgarriff and Rosenzweig (2000).

Many sense inventories have been taken from traditional paper-based (and from machine-readable) dictionaries. They benefit from having been developed by human experts – lexicographers – but are often difficult to exploit computationally. First, the sense distinctions are designed for the application of helping human users understand the meaning of a word in context, not to elucidate how the senses are different. Secondly, while dictionary-ese is very precise and formal compared to most human languages, it is not formal enough.

WordNet has also been a popular choice, because, first, it is designed as a research artefact; secondly, it is freely available on terms that do not constrain researchers (as is not usually the case with paper-based dictionaries); and thirdly, it is already so widely used that it approaches the status of a *de facto* standard for English and other languages where WordNets are available. Although WordNet has been criticized for lack of lexicographic rigour, and for its thesaurus-design (focusing on similarities between different words) rather than dictionary-design (focusing attention on the different meanings of the same word), the pro-arguments retain a great force.

2.4 Sense-tagged corpora

The DARPA evaluation methodology is to score systems by measuring their output against output generated by people, and a substantial manually annotated gold standard corpus is required. High inter-annotator agreement and replicability are necessary, or the gold standard is fool's gold. Low inter-annotator agreement, assuming qualified annotators, indicates that the sense inventory is inadequate, or that the task is too difficult or ill-defined; see Calzolari *et al.* (this issue). Table 1 lists some major sense-annotated corpora currently available.

3 History of WSD evaluation

Gale, Church and Yarowsky (1992a) review, exhaustively and somewhat bleakly, the state of affairs up to 1992, several years before a DARPA model for WSD evaluation was adopted. 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 compared the performance of one of their systems (Yarowsky 1992) to that of other WSD systems for which accuracy figures were available (considering

Table 1. *Major sense-annotated corpora*

line, hard and *serve* corpora
3 lemmas, 12,000+ instances. Inventory: Wordnet 1.5. Text sources: Wall Street Journal, American Printing House for the Blind, San José Mercury. Leacock, Towell, and Voorhees (1993), Leacock, Chodorow and Miller (1998). http://www.d.umn.edu/∼tpederse/data.html

interest corpus[a]
1 lemma, 2,369 instances. Inventory: LDOCE Text sources: Wall Street Journal. Rebecca Bruce and Jan Wiebe. http://crl.nmsu.edu/cgi-bin/Tools/CLR/clrcat#I9

Hector[b]
ca 300 lemmas, 200,000 instances. Inventory: Hector. Text source: A 20M-word pilot for the British National Corpus. Atkins (1993).

Semcor
23,346 lemmas, 234,113 instances. Inventory: WordNet 1.5, 1.6. Text sources: 80% Brown corpus, 20% a novel, *The Red Badge of Courage*. Fellbaum (1998). http://cogsci.princeton.edu/∼wn

DSO Corpus
191 lemmas, 192,800 instances. Inventory: WordNet 1.5. Text sources: Brown Corpus, Wall Street Journal. Ng and Lee (1996).

Open Mind Word Expert
180 lemmas, 55,000[c] instances. Inventory: WordNet 1.6, 1.7. Text source: Penn treebank, LA Times, others. Chklovski and Mihalcea (2002). http://www.teach-computers.org/word-expert.html

HKUST-Chinese
38,725 sentences. Inventory: Hownet. Text source: Sinica corpus. Gan Kwok-Wee and Wong Ping-Wai. http://godel.iis.sinica.edu.tw/CKIP/hk/index.html http://www.keenage.com

Swedish corpus
179,151 instances tagged. Inventory: Gothenburg lexical database. Text source: The SUC Corpus. Jaerborg, Kokkinakis, and Toporowska (2002).

Image captions
2,304 lemmas, 8,816 instances. Inventory: WordNet 1.5. Text source: Image captions of an image collection. Smeaton and Quigley (1996). http://www.computing.dcu.ie/∼asmeaton/SIGIR96-captions/

Senseval-1
See table 2. Kilgarriff and Palmer (2000).

Senseval-2
See table 3. Edmonds and Cotton (2001).

[a] Instances of 11 other words have been tagged on a similar basis but the data has not been made available.

[b] Hector is an Oxford University Press and DEC dictionary research project.

[c] Includes duplicates but the number is growing daily; this is an on-line resource that web-users can add to at any time.

each word individually). 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)

To remedy this state of affairs, Gale, Church and Yarowsky introduced baselines (i.e. methods to compute upper and lower bounds on performance) which could be used to assess both the relative difficulty of disambiguating different words and the relative performance of different WSD systems. They estimated lower and upper bounds of 75% (by always choosing the most frequent sense in the test set) and 96.8% (by performing a sense discrimination experiment with human judges), respectively.

In 1993, Leacock, Towell and Voorhees reported on a sense-tagged corpus built by annotating the occurrences of the word *line*, which they used to compare three different WSD algorithms. Mooney (1996) used the same corpus to compare seven machine learning algorithms with different biases.

Of course, the concern is not just with evaluating different algorithms on the same corpus, but also the same algorithm on different corpora. Ng (1997) investigated differences in using different corpora for training an algorithm.

The topic was raised as the central issue of an ACL-SIGLEX workshop in Washington, April 1997. Resnik and Yarowsky (1999) made some practical proposals for an evaluation exercise which were enthusiastically welcomed and opened an energetic debate. From this debate was born Senseval.

4 Sensevals past, present and future

4.1 Senseval-1

Senseval-1 was held in 1998 (Kilgarriff and Palmer 2000). A lexical-sample methodology for the large-scale evaluation of WSD systems was agreed, whereby

- a sense inventory is chosen
- a stratified random sample is taken from the lexicon, with sub-samples for part of speech, frequency band, and number of senses
- corpus instances of a few sentences (or more) around the target words in a large corpus are selected for each target word
- the target word in each corpus instance is tagged by at least two human judges
- the tagged corpus is divided into a training and test corpora
- participants train (if supervised) and run their systems on the test corpus *and*
- the system answers are scored against the held back tags of the test corpus.

Tasks were designed for English, French, and Italian. Table 2 gives some statistics about the corpora, the participants, and the scores of the best performing systems.

Table 2. *Results of* Senseval-1*, tabulated from Kilgarriff and Palmer (2000)*

Language	Systems	Lemmas	Instances[a]	IAA[b]	Baseline[c]	Best score
French	4	60	3000	~0.70	0.53/0.40/0.22[d]	0.60/0.63/0.71
Italian	2	58	2701	~0.78	–[e]	–
English	17	41	8448	0.95	0.57	0.78

[a] Total instances annotated in both training and test corpora.
[b] Inter-annotator agreement for French and Italian is average pairwise agreement; for English it is based on a replicability experiment (Kilgarriff 1999).
[c] Generally, choosing the corpus-attested most frequent sense.
[d] Scores for adjectives, nouns, and verbs, respectively.
[e] Data not available.

It was promising that systems could achieve 78% accuracy on the English task over a broad selection of 41 words.[1]

Senseval-1 produced a set of benchmarks for WSD system performance. It set out to establish the viability of WSD as a separately evaluable NLP task. This was conclusively proven: the replicability, and thus validity, of the gold-standard corpus (for English Senseval-1) was 95% (Kilgarriff 1999).

At the Senseval-1 workshop, in Herstmonceux, Sussex, UK, and afterwards, the way forward was discussed extensively. While the evaluation had worked well, and it was desirable to repeat some tasks (which, amongst other things, would permit the measurement of progress), there were some aspects of WSD which were not covered by the task design. Future Sensevals needed a wider range of tasks, including ones where WSD was contributing to an application (machine translation or information retrieval). It was also desirable to have tasks for a wider range of languages.

4.2 Senseval-2

Senseval-2 was organized in 2000–2001. Its goals were to encourage tasks in new languages, to encourage more participants to enter their systems, and to broaden the range of tasks. It was successful: Senseval-2 evaluated WSD systems on three tasks on 12 languages as shown in Table 3. The lexical sample task is designed as above. The translation task (Japanese only) is a lexical sample task in which word sense is defined according to translation distinction. An all-words task, a recommendation following Senseval-1, is a task in which all of the content words in a portion of running text are to be tagged. Senseval-2 scored 94 systems submitted by 35 different research teams.

Table 3 also gives the accuracy of the best performing system on each task. Note that for English, the scores are much lower than for Senseval-1. Kilgarriff

[1] Systems were scored in terms of precision (percentage of right answers in the set of answered instances), recall (percentage of right answers over all instances), and coverage (percentage of instances attempted). Most systems attempted all instances, giving a recall figure equal to the precision figure. Precision figures for systems attempting all instances are reported in Table 2 (and Table 3).

Table 3. *Results of* Senseval-2, *tabulated from Edmonds and Cotton (2001)*

Language	Task[a]	Systems	Lemmas	Instances[b]	IAA[c]	Baseline[d]	Best score
Czech	AW	1	–[e]	277,986	–	–	0.94
Basque	LS	3	40	5,284	0.75	0.65	0.76
Dutch[f]	AW	1	1,168	16,686	–	0.75	0.84
English	AW	21	1,082	2,473	0.75	0.57	0.69
English	LS	26	73	12,939	0.86	0.48/0.16[g]	0.64/0.40
Estonian	AW	2	4,608	11,504	0.72	0.85	0.67
Italian	LS	2	83	3,900	0.21	–	0.39
Japanese	LS	7	100	10,000	0.86	0.72	0.78
Japanese	TL	9	40	1,200	0.81	0.37	0.79
Korean	LS	2	11	1,733	–	0.71	0.74
Spanish	LS	12	39	6,705	0.64	0.48	0.65
Swedish	LS	8	40	10,241	0.95	–	0.70

[a] AW: all-words task, LS: lexical sample, TL: translation memory.

[b] Total instances annotated in both training and test corpora. In the default case, they were split 2:1 between training and test sets.

[c] Inter-annotator agreement is generally the average percentage of cases where two (or more) annotators agree, before adjudication. However, there are various ways in which it can be calculated, so the figures in the table are not all directly comparable.

[d] Generally, choosing the corpus-attested most frequent sense, although this was not always possible or straightforward.

[e] A dash '–' indicates the data was unavailable.

[f] The Dutch task was not run during Senseval-2, however the data was prepared for Senseval-2.

[g] Supervised and unsupervised scores are separated by a slash.

(2002) suggests the cause is WordNet, implying that WordNet's sense distinctions are less clear and less motivated than hector's, as used in Senseval-1. Trang Dang, Palmer and Fellbaum (2002), however, present evidence that the words chosen for Senseval-2 were in fact more difficult to disambiguate.

Senseval has created datasets of substantial value to the community.[2] But, like the data and samples brought back from even a short archaeological expedition to Egypt, it will take a long time to analyse. Researchers are only beginning to uncover differences between systems (related to the methods, the knowledge sources, and the contextual features used), to determine the difficulty classes of words to be disambiguated, and to analyse the sense distinctions made in the lexicon. The papers in this issue begin this analysis, but see also Stevenson and Wilks (2001) and the papers in Edmonds, Mihalcea and Saint-Dizier (2002).

4.3 A typology of evaluations

In Table 4 we propose a breakdown of evaluation types based on the basic set up (*in vitro* or *in vivo*) and the type of sense inventory used. *In vitro* or 'glass box'

[2] Senseval-1 and Senseval-2 data sets and the results of all Senseval-2 systems are available at http://www.sle.sharp.co.uk/senseval2.

Table 4. *Evaluation set up versus sense inventory*

Sense inventory	*In vitro* evaluation	*In vivo* evaluation
Explicit application-independent	SENSEVAL	?
Explicit, defined by an application or domain	E.g. senses as translation equivalents	E.g. improvement in machine translation or information extraction as the task
Implicit, defined by application in a domain	E.g. senses as word or context clusters	E.g. improvement of information retrieval as the task

evaluation is evaluation outside of any particular application. It allows for more generic WSD systems, which can be more easily analysed and compared in detail. *In vivo* evaluation is done within the context of a particular application, which results in a more realistic assessment of a system's ultimate utility.

An explicit sense inventory is written down in a form external to an application; examples include application-independent resources such as generic monolingual and bilingual dictionaries and WordNet, and application-specific resources such as translation lexicons. In contrast, implicit inventories are not produced by, or designed to be shown to, a person. They may not even exist as explicit objects in the system. An example is the word clusters formed through document comparisons during information retrieval.

Note that SENSEVAL, with the honourable exception of the Japanese translation task, occupies to date only one cell in the table. No significant comparative evaluation has taken place in the other cells. The plan for SENSEVAL-3 is to move into these cells, in particular, *in vitro* evaluation with an explicit, application-defined sense inventory.

5 The papers

This special issue grew out of the SENSEVAL-2 workshop. Papers were invited on any topic in WSD related to evaluation, with emphasis on analysis of SENSEVAL-2 data and results.

The first three papers evaluate supervised Machine Learning (ML) approaches.

Yarowsky and Florian present what is probably the most comprehensive study to date of the effect that various data characteristics (such as sense granularity, sense entropy, feature classes, and context 'window' size) have on the performance of a range of supervised approaches. They show that a distinction between discriminative and aggregative algorithms is empirically motivated. A discriminative algorithm relies on the best feature (or few features) in the context to make its 'winner-takes-all' decision, whereas an aggregative algorithm integrates all available evidence for each sense to make a decision. For some words, discriminative is best, for others, aggregative is. They also find that the algorithm itself has a significantly lower impact on performance than the feature space within which the algorithm operates. Their algorithms were tested on lexical sample tasks for Basque, English, Spanish and Swedish.

In counterpoint, Hoste *et al.* show that the parameter settings as well as the feature space of an algorithm have a great impact on performance. They explore the parameter space for individual word experts, using memory-based learning algorithms and discover that results vary wildly according to the parameters selected, and that generalizations about 'good' paramater settings for classes of problem were mostly invalid. Their system is tested on English and Dutch all-words tasks.

Hoste *et al.*'s finding is rather far-reaching and alarming. Until now, most researchers assumed that it was reasonable to use ML packages with default settings, and that, on applying an ML package to a data set, one will get results that tell us about the performance of that algorithm on that data set. Sadly, this is shown to be false. One algorithm may perform significantly better than another with one set of parameters, and significantly worse with another. The finding has parallels with recent work by Banko and Brill (2001), in which they establish that it is hazardous to declare one ML algorithm superior to another for a given task on quite different grounds. They experimented with a million-word training corpus, and then with a billion-word one. Across the board, performance improved with increased training data, but more so for some algorithms (with some parameter settings, one should add) than for others, so what had seemed to be the best algorithm for the task, as long as training corpus size had not been seen as a parameter needing investigation, no longer was. This research does an impressive job of undermining what we might have thought we were learning about ML algorithms in language technology!

Given that different ML approaches have different strengths and weaknesses on the same data set – the well-known problem of bias in machine learning – can we take the best of all worlds? Florian *et al.* demonstrate that we can. First, as expected, their experiments show that there is significant variation *and* inter-agreement in word-sense classifier performance across different languages and data sizes. Then, they analyse several techniques for classifier combination, including count-based voting, confidence-based voting, and probability mixture models. Meta-voting, in which all of the combined classfiers vote, achieves an accuracy higher than any SENSEVAL-2 system.

The next two articles explore new techniques for WSD, the first by semi-supervised all-words WSD, the second by exploiting domain knowledge.

Mihalcea describes the high-performing SMU system, which participated in both the English-all-words task (SMUaw) and English-lexical-sample task (SMUls). It uses a combination of pattern learning – an extended model of n-grams that surround the target words – and active feature selection whereby a subset of features for each target word is first learned before applying the main learning algorithm. The system performs well on the all-words task apparently because the learned patterns benefit from a 160,000-word sense-tagged corpus automatically built through heuristics, in an interesting use of bootstrapping. In contrast, on the lexical sample task, where training data was available, active feature selection appears to be the key. Mihalcea shows that the same feature can either improve or degrade performance depending on the target word.

Magnini *et al.* show that using domain information can improve performance on WSD. The authors first explore a 'one domain per discourse' hypothesis, which backs up a claim that a text often has a prevalent domain (a close relative to Gale, Church and Yarowsky's (1992b) 'one sense per discourse'). They find that domain is not a fixed notion for a text; several domains can be represented at once and the domain can vary throughout a text. Moreover, the domain relevant to disambiguating each word in a text varies over the text. The authors evaluate a method of WSD that determines the relevant domain for each disambiguation decision. This view of lexical ambiguity assumes that a word has different senses in different domains, and once the domain is identified, the word is no longer ambiguous. The method obtains good results on the English tasks in Senseval-2 without using any other features.

Now, the Senseval model can be seen as biased towards a 'linguistic' understanding of sense ambiguity, in contrast to a 'domain' one. In the linguistic understanding, one expects to exploit evidence for disambiguation within a relatively small and narrowly defined context, e.g. in the sentence or the page. This is the only kind of approach which Senseval has, to date, been able to assess. In vivid contrast, the great bulk of sense 'disambiguation' undertaken in commercial MT takes a domain view. MT systems have different lexicons for different domains, and the lexicon is chosen by the system user who knows what domain he or she is working in. No disambiguation process is required. This perspective on WSD has recently been espoused by Vossen (2001) and Buitelaar and Sacaleanu (2001).

One possible synthesis of the two opposing perspectives is to view disambiguation as a two-stage process, where the first stage is domain identification from some fixed list of domains which are also marked up in the lexicon. Magnini *et al.* take us along this route.

In the final paper, Calzolari *et al.* pull back from particular algorithms and discuss the evaluation methodology. They consider how the quality of the lexical-semantic resources can affect the evaluation of WSD systems, and in reverse, how Senseval can be used to evaluate the quality of the resources. Their study is based on the Italian tasks in Senseval-1 and Senseval-2. They conclude that traditional, dictionary-based, resources are inadequate because of the divide they create between their own decontextualised nature and the contextualised nature of word sense disambiguation. They propose feeding back from manual and automatic tagging into the construction of new resources. They suggest the way forward is to design a dynamic lexicon that can cope with the complex relationships between lexicon and corpus.

6 Conclusion

It is clear from the research presented herein that we have not reached the limit of WSD system performance – systems do not yet challenge the upper bounds on the task. Machine learning approaches will be the key to success in the near future.

Current work in bootstrapping, co-training, choosing feature-spaces, and parameter estimation promises better performance, in the face of the lack of data. But how far can such methods take us? We can't really say until we train and run the systems on much larger annotated corpora, which are coming as a result of SENSEVAL.

We also think application-specific and domain-specific WSD systems will focus research in the coming years. The problems of real applications are more concrete, so the results of common evaluations will be easier to assess regarding system performance and utility. SENSEVAL is moving in this direction and will define new application-specific tasks for WSD.

We are just starting to understand the intricacies of lexical ambiguity and lexical semantics. While it might appear that SENSEVAL is just about improving performance on WSD, its underlying mission is to develop our understanding of the lexicon in particular, and language in general.

Acknowledgments

Many good papers were submitted to this special issue, but we could choose only these few. We would like to thank all the authors who submitted papers, for they keep this area of research vibrant. We also thank the reviewing committee who helped us to decide which papers to accept and provided many excellent reviews for the authors. The reviewers were:

Eneko Agirre (University of the Basque Country), Roberto Basili (University of Roma Tor Vergata), Nicoletta Calzolari (Istituto di Linguistica Computazionale del CNR, Pisa), Key-Sun Choi (Korean Advanced Institute for Science and Technology), Christiane Fellbaum (Princeton University), Chu-Ren Huang (Academia Sinica, Taiwan), Nancy Ide (Vassar College), Sadao Kurohashi (Tokyo University), Martha Palmer (University of Pennsylvania), Philip Resnik (University of Maryland), German Rigau (Universitat Politècnica de Catalunya), Antal van den Bosch (Tilburg University), Jean Véronis (CILSH, Université de Provence), Clare Voss (Army Research Laboratory, USA), Piek Vossen (Irion Technologies BV), Pete Whitelock (Sharp Laboratories of Europe), Janyce Wiebe (University of Pittsburgh), David Yarowsky (Johns Hopkins University), Antonio Zampolli (Istituto di Linguistica Computazionale del CNR, Pisa).

Kilgarriff's involvement was supported by the UK EPSRC under grants WASPS (GR/M54971) and MATS (GR/R02337/01).

We are very grateful to all those who responded to our query to the CORPORA mailing list in August 2002 regarding the existence and availability of sense-annotated resources. We also thank Antal van den Bosch and Ken Church for comments on previous drafts.

References

Atkins, S. (1993) Tools for computer-aided lexicography: The Hector project. *Papers in Computational Lexicography: COMPLEX '93*, Budapest.

Banko, M. and Brill, E. (2001) Scaling to very very large corpora for natural language disambiguation. *Proceedings 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Toulouse.

Buitelaar, P. and Sacaleanu, B. (2001) Ranking and selecting synsets by domain relevance. *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*. Pittsburgh.

Chklovski, T. and Mihalcea, R. (2002) Building a sense tagged corpus with open mind word expert. *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pp. 116–122. Philadelphia.

Edmonds, P. and Cotton, S. (2001) Senseval-2: Overview. *Proceedings SENSEVAL-2: The Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pp. 1–5. Toulouse.

Edmonds, P., Mihalcea, R. and Saint-Dizier, P., eds. (2002) *Proceedings ACL*-02 *Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*. Philadelphia.

Fellbaum, C., ed. (1998) *WordNet: An Electronic Lexical Database*. MIT Press.

Gale, W. A., Church, K. W. and Yarowsky, D. (1992a) Estimating upper and lower bounds on the performance of word-sense disambiguation programs. *Proceedings 30th Annual Meeting of the Association for Computational Linguistics*, pp. 249–256.

Gale, W. A., Church, K. W. and Yarowsky, D. (1992b) One sense per discourse. *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 233–237. New York.

Hanks, P. (2000) Do word meanings exist? *Computers in the Humanities* **34**(1–2).

Ide, N. and Véronis, J. (1998) Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics* **24**(1): 115–141.

Jaerborg, J., Kokkinakis, D. and Toporowska Gronostaj, M. (2002) Lexical and textual resources for sense recognition and description. *Proceedings LREC 2002*. Las Palmas, Spain.

Jurafsky, D. and Martin, J. (2000) *Speech and Language Processing*. Prentice Hall.

Kilgarriff, A. (1997) "I don't believe in word senses". *Computers in the Humanities* **31**(2): 91–113.

Kilgarriff, A. (1999) 95% replicability for manual word sense tagging. *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 277–278. Bergen.

Kilgarriff, A. (2001) Generative lexicon meets corpus data: The case of non-standard word uses. In: Bouillon, P. and Busa, F., editors, *The Language of Word Meaning*, pp. 312–328. Cambridge University Press.

Kilgarriff, A. (2002) English lexical sample task description. *Proceedings SENSEVAL-2: The Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pp. 17–20. Toulouse.

Kilgarriff, A. and Palmer, M. (2000) Introduction to the special issue on SENSEVAL. *Computers in the Humanities* **34**(1–2): 1–13.

Kilgarriff, A. and Rosenzweig, J. (2000) Framework and results for English SENSEVAL. *Computers in the Humanities* **34**(1–2): 15–48.

Leacock, C., Chodorow, M. and Miller, G. (1998) Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics* **24**(1): 147–166.

Leacock, C., Towell, G. and Voorhees, E. (1993) Corpus-based statistical sense resolution. *Proceedings ARPA Human Language Technology Workshop*. Morgan Kaufman.

Manning, C. and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. MIT Press.

Mooney, R. J. (1996) Comparative experiments in disambiguating word senses: An illustration of the role of bias in machine learning. *Proceedings of the First Conference on Empirical Methods in Natural Language Processing*.

Ng, H. T. (1997) Getting serious about word sense disambiguation. *Proceedings of the Workshop on Tagging Text with Lexical Semantics: What, Why, and How?*, pp. 1–7.

Ng, H. T. and Lee, H. B. (1996) Integrating multiple sources to disambiguate word sense: An exemplar-based approach. *Proceedings 34th Annual Meeting of the Association for Computational Linguistics*.

Pustejovsky, J. (1995) *The Generative Lexicon*. MIT Press.

Resnik, P. and Yarowsky, D. (1999) Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *J. Natural Lang. Eng.* **5**(2): 113–134.

Smeaton, A. F. and Quigley, I. (1996) Experiments on using semantic distances between words in image caption retrieval. *Proceedings of the 19th International Conference on Research and Development in Information Retrieval* (*SIGIR96*), pp. 174–180. Zurich, Sweden.

Stevenson, M. and Wilks, Y. (2001) The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics* **27**(3): 321–349.

Trang Dang, H., Palmer, M. and Fellbaum, C. (2002) Making fine-grained and coarse-grained sense distinctions, both manually and automatically. Submitted.

Vossen, P. (2001) Extending, trimming and fusing WordNet for technical documents. *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*. Pittsburgh.

Yarowsky, D. (1992) Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. *Proceedings of the 14th International Conference on Computational Linguistics* (*COLING-92*). Nantes.