University of Brighton

ITRI-00-21

Introduction, Special Issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs

Adam Kilgarriff and Martha Palmer

December, 2000

Also published in Computers and the Humanities 34:1-2, pp. 1-13.

Supported by the EPSRC under Grants K18931 and M54971

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

Introduction to the Special Issue on SENSEVAL

A. Kilgarriff
ITRI, University of Brighton

M. Palmer
University of Pennsylvania

(Received ..... ; Accepted in final form .....)

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 Lexicons Special Interest Group of the Association for Computational Linguistics), 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

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* or *médicament*, 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 extractions systems may make wrong assertions; text-to-speech systems will confuse 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 (IV98) and the reader is directed to that paper for historical background and kinds of methods that have been used.

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 ARPA competitive evaluations for speech recognition, dialogue systems, information retrieval and information extraction has been that the focus provided by an evaluation serves to bring research communities together, forces consensus on what is critical about the field, and leads to the development of common resources, all of which then stimulates further rapid progress (see, eg. (Gai98)).

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.

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 systems 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.

(GCY92) reviews, exhaustively and somewhat bleakly, the state of affairs at that time. They open:

> 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' (Yar92) 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 ARPA 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 ARPA 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 non-controversial level of semantic representation that might be more amenable to evaluation than other more problematic levels.

Resnik and Yarowsky, (RY97), made some practical proposals for evaluation of WSD systems using machine learning techniques which 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, that is, can they be replicated reliably?

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 success of the workshop gives us an unequivocal yes to our first two question, but answering the third one is not surprisingly much more complicated.

3. Papers

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 to be hegemonic, 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,

[1] The name is due to David Yarowsky.

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 through the whole process.

An evaluation needs a scoring metric, and one of the issues raised by (RY97) was that the simple metric, whereby a correct response scores 1 and anything else scores 0, was 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 Issue which is not specific to WSD, Melamed and Resnik present a scoring scheme meeting the desiderata. The scheme underlay the scoring strategy 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.

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 3 shows, for each task, how many participating systems, research groups and Special Issue papers there are.[2]

[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.

Table I.

	Systems	Research Groups	Papers	Brief note
English	18	17	15	2
French	5	4	1	3
Italian	2	2	1	0
TOTALS	25	23	17	5

For most of the 25 participating systems, there is a paper in the Special Issue (and for five 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 seemed to fall into two opposed 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. The "unsupervised systems" — those relying on previously created lexical resources — made extensive use of the training data to fine tune their system, or the "supervised systems" — those relying on machine learning and 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 machine-learning approaches 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, and collocates at specific positions. 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. interesting as their Bayesian model did assume independence. One system (Hawkins's) used some manually rather than automatically derived features, with the manual acquisi-

tion organised so that it bootstrapped from untagged training material and took little time. Veenstra et al. improved their system performance when they optimised the settings in their model for each individual word based on performance in a 10-fold cross validation. They did not get the same settings appearing time and again for several words, but rather quite distinct settings for each individual lexical item. Approaches that are sensitive to such individual differences are clearly necessary, but the amount of training data is disconcerting.

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.

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 arguments 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 by 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; for the case for it, see section 2 of Kilgarriff and Rosenzweig. 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) takes this opportunity to respond. SENSEVAL is, from one point of view, an experiment designed to replace scepticism, about both the reality of words 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%); at what level do state-of-the-art systems perform (75–80%) (both answers relative to a fine-grained, corpus-based dictionary). 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 (Kil99).

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

4. Responses to criticisms

Firstly, there are some recurring criticisms of the ARPA quantitative evaluation paradigm, (SOH99). These are

- 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.
- There is a substantial overhead (financial, temporal and emotional) involved both in setting up the evaluations and in participating in them.
- It encourages a competitive (as opposed to collaborative) ethos ;
- 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). The second does not apply to this first, small-scale evaluation (where much was done on goodwill) but is likely to apply for future, hopefully large-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 (Hir98) 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 ARPA 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, in stating reasons for not participating, stated that, for them, WSD occurred as a byproduct of deeper reasoning so it would not make sense to participate in an exercise that treated WSD outputs as of interest in their 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 (Kil97)).

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 (**?**; SJG96; Gai98). 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, "is WSD a separable NLP subtask?". It would seem some parts of the task, such as homograph resolution, are, while others, such as metaphor resolution, are not. 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.

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 (**?**).

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 work to use sense inventories which were appropriate for Machine Translation tasks. The foundational work of Resnik and Yarowsky, (RY97), (RY99) and Ide (this volume) on clustering together monolinugual usages based on similar translations provides the 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, and also that different translations in another language do not always correlate with a valid sense distinction in the source language. 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 sense distinctions, and, given our current state of knowledge, any such evidence is to be embraced.

The Herstmonceux SENSEVAL will be the first of a series. We will continue to investigate the correlations between human performance and machine performance, we will take as much advantage as possible of empirical methods in data preparation and system evaluation, and we welcome suggestions and comments on both our methodology and results from the community.

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.

This Special issue contains detailed accounts of how the evaluation exercises were set up, and the results. Most of the participating systems are described and there are positions 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 special issue, 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 Asociation), 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`

References

Robert Gaizauskas. Evaluation in language and speech technology: Introduction to the special issue. *Computer Speech and Language*, 12(4):249–262, 1998.

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

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

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

Adam Kilgarriff. Foreground and background lexicons and word sense disambiguation for information extraction. In *Proc. Workshop on Lexicon Driven Information Extraction*, pages 51–62, Frascati, Italy, July 1997.

Adam Kilgarriff. Generative lexicon meets corpus data: the case of non-standard word uses. In Pierrette Bouillon and Frederica Busa, editors, *Word Meaning and Creativity*, chapter Cambridge. Cambridge University Press, forthcoming, 1999.

Philip Resnik and David Yarowsky. 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 1997. SIGLEX (Lexicon Special Interest Group) of the ACL.

Philip Resnik and David Yarowsky. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering Journal*, to appear, 1999.

Karen Sparck Jones and Julia Galliers. *Evaluating Natural Language Processign Systems: An Analysis and Review.* Springer-Verlag, Berlin, 1996.

R. Sproat, M. Ostendorf, and A. Hunt. The need for increased speech synthesis research. Report of the 1998 NSF Workshop for Discussing Research Priorities and Evaluation Strategies in Speech Synthesis, March 1999.

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

Address for correspondence: ITRI, University of Brighton, Lewes Road, Brighton BN2 4GJ, UK