Evaluating Automatic Semantic Taggers

Philip Resnik
Dept. of Linguistics/UMIACS
University of Maryland
College Park, MD 20742
resnik@umiacs.umd.edu

David Yarowsky
Dept. of Computer Science/CLSP
Johns Hopkins University
Baltimore, MD 21218
yarowsky@cs.jhu.edu

Unlike the problems of part-of-speech tagging and parsing, where commonly utilized training and test sets such as the Brown Corpus and Penn Treebank have existed for a number of years, evaluation of word sense disambiguation sytems is not yet standardized. In fact, most previous work in sense disambiguation has tended to use different sets of polysemous words, different sense inventories, different evaluation metrics and different test corpora. This working session will address these problems and seek solutions to them. Examples of issues for discussion include:

- How should part-of-speech-level distinctions be treated when evaluating WSD systems?
- How should sense inventories be defined so as not to be biased in favor of certain disambiguation methods, such as those based on selectional restriction, topic codes, hierarchical ontologies, or aligned multilingual corpora? Or are such biases ok?
- What evaluation metrics are appropriate for the WSD task?
- What characteristics should common test suites exhibit? How and by whom should they be developed?
- Would a MUC-style competitive evaluation program be beneficial or detrimental to progress in the WSD field?
- What special problems exist when evaluating WSD performance on verbs?
- What special problems exist when evaluating WSD performance in a multi-lingual setting?
- What additional issues arise in evaluating more complex semantic tagging, going beyond sense disambiguation as traditionally defined?
- How should regular polysemy and metaphor be treated in WSD evaluation?
- Can a common evaluation framework satisfy the needs and limitations of both supervised and unsupervised sense disambiguation methods?

References

C. Leacock, G. Towell and E. Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings, ARPA Human Language Technology Workshop*, pp. 260–265, Plainsboro, NJ.

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

P. Resnik and D. Yarowsky 1997. A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, DC.

Y. Wilks. and M. Stevenson 1996. The grammar of sense: Is word-sense tagging much more than part-of-speech tagging? cmp-lg/9607028.