Putting Meaning into Your Trees

Martha Palmer
University of Pennsylvania
mpalmer@cis.upenn.edu

The meaning of a sentence is an essential aspect of natural language understanding, yet an elusive one, since there is no accepted methodology for determining it. There is not even a consensus on criteria for distinguishing word senses. Clearly a more robust technology is needed that uses data-driven techniques. These techniques typically rely on supervised machine learning, so a critical goal is the definition of a level of semantic representation (sense tags and semantic role labels) that could be consistently annotated on a large scale. We have been training automatic WSD systems on the English sense-tagged training data based on WordNet that we supplied to SENSEVAL2 (Dang & Palmer, 2002). A pervasive problem with sense tagging is finding a sense inventory with clear criteria for sense distinctions. WordNet is often criticized for its subtle and fine-grained sense distinctions. Perhaps more consistent and coarse-grained sense distinctions would be more suitable for natural language processing applications. Grouping the highly polysemous verb senses in WordNet (on average reducing the >16 senses per verb to 8) provides an important first step a more flexible granularity for WordNet senses that improves both inter-annotator agreement (71% to 82%) and system performance (60.2% to 69%) (Dang & Palmer, 2002). The Frameset sense tags associated with the PropBank, as discussed below, provide an even more coarse-grained and easily replicable level of sense distinctions.

Based on a consensus of colleagues participating in the ACE (Automatic Content Extraction) program, we have developed a Proposition Bank, or PropBank, which provides semantic role labels for verbs and participial modifiers for the 1M word Penn Treebank II corpus (Marcus, 1994). VerbNet classes have proved invaluable for defining the appropriate semantic roles in this endeavor (Dang, et. al., 1998). For example, *John* is the Agent or Arg0 of *John broke the window*, *IBM* is the Theme or Arg1 of *IBM rose 1.2 points*. In addition, for just over 700 of the most polysemous verbs in the Penn TreeBank, we have defined two or more Framesets – major sense distinctions based on differing sets of semantic roles (Palmer, et al, submitted). These Framesets overlap closely (95%) with our manual groupings of the SENSEVAL2 verb senses, and thus they can be combined to provide an hierarchical set of sense distinctions. The PropBank is complete and a beta-release version was made publicly available through LDC in February for use in the CoNLL-04 shared task. There is a complementary lexicography project at Berkeley, Chuck Fillmore's FRAMENET, which provides representative annotated samples rather than broad-coverage annotation, and there are current plans to combine these resources and train automatic labelers for English and Chinese. The automatic semantic role labelers we are building use features that are very similar to our WSD system features, and we find that semantic role label features improve WSD while sense tag features improve semantic role labeling (Gildea & Palmer, 2002).

References

Dang, H.T., Kipper, K., Palmer, M., Rosenzweig, J. (1998) Investigating regular sense extensions based on intersective Levin classes. *Coling/ACL-98, pp 293-300,* Montreal CA, August 11-17, 1998.

Dang, H. T. and Palmer, M., (2002). Combining Contextual Features for Word Sense Disambiguation. In *Proceedings of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, Pa.

Gildea, D and Palmer, M., (2002) The Necessity of Parsing for Predicate-Argument Recognition, *ACL-02*, Philadelphia, PA, July 7-12.

Marcus, M, (1994), The Penn TreeBank: A revised corpus design for extracting predicate argument structure, In *Proceedings of the ARPA Human Language Technology Workshop*, Princeton, NJ.

Palmer, M., Gildea, D., Kingsbury, P. (submitted) The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics.*