Mapping Low-level Linguistic Representations onto Interpretations for Cognitive Systems using Data Driven Methods

Mark Steedman, Edinburgh

There has been some recent success in parsing with wide coverage (Collins and Singer 1999; Charniak 2000; Bod 2001. The most successful techniques are based on rule-based supervised methods—that is, the use of automatically induced rule-based grammars and head-dependency-based statistical models, both derived from human annotated data such as the Penn Wall Street Journal Treebank. Such parsers perform much better than hand built grammars, and much better than grammars built using unsupervised methods—that is, induction of grammars and models from unannotated text. Such treebank parsers have proved useful—up to a point—for open-domain question answering (Moldovan et al. 2000), Grammar-assisted language modeling for ASR (Charniak 2001), and many other practical applications. It has also been possible to extend the relevant techniques to parsers for more expressive grammars recovering "semantic" dependencies relevant for to building interpretable structure.

However, such parsers are subject to some serious shortcomings, which severely limit such applications. The grammars they support tend to overgenerate, often producing uninterpretable structures. They also to undergenerate, failing to find any analysis at all. This does not seem to be a problem that will be solved by Moore's law. All such parsers seem to be approaching a similar overall asymptotic performance level, whatever measure is adopted and whatever grammar is involved. The source of this weakness lies in both the nature of the grammatical and statistical models and the paucity of the data from which they are derived.

The first of these is probably quite easy to remedy. It is already clear that it is possible to extend the treebank techniques to parsers for more expressive grammars recovering the "semantic" dependencies that are needed to build interpretable structure. However, while such a move will improve the quality of interpretations for some classes of sentence, it will not make much impact on overall performance, since such constructions are themselves relatively sparse. The real bottleneck is that the one million words of the Penn Treebank, the largest and most accurate treebank that we have, at least for English, is not nearly enough to show all words in all possible constructions, or to provide a reliable head dependency model.

For example, the Penn treebank includes very few questions, and none at all of the form "Which X is Y?". This means that performance in parsing, say, the questions for the TREC question answering competition is startlingly bad if these parsers are used naively.

To extend the treebank training data with material for new genres in prohibitively expensive. It alos will not in general sove the problem: Zipf's law tells us that we would need an order of magnitude more data to achieve any noticeable improvement—and that that still would not be enough data. We must turn to other techniques for extensing the supervised-learning based parsers.

While handbuilt rersources such as machine-readable dictionaries might offer an easy way to do this, in practice they too seem to miss too much of the necessary data. There is probably no alternative to investigating hibrid techniques that improve these parsers using unsupervised learning over large amounts of raw text.

Key technologies that have ben investigated include co-training; clustering and principal components analysis over raw text, using features from treebank grammars as guides; high pre-

cision techniques involving high frequency function word n-grams and criteria such as minimum description length. None of these techniques has done particularly well in isolation: the grand challenge is to adapt them to piggy-back on treebank grammars, on the assumption that every con struction probably has been seen in a million words, but not every word-construction pair

References

Bod, Rens. 2001. "What is the Minimal Set of Fragments that Achieves Maximal Par se Accuracy?" In *Proceedings of the 39th Meeting of the ACL*. Toulouse, France.

Charniak, Eugene. 2000. "A Maximum-Entropy-Inspired Parser." In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 132–139. Seattle, WA.

Charniak, Eugene. 2001. "Immediate-Head Parsing for Language Models." In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse*, 116–123. San Francisco, California: Morgan Kaufmann Publishers.

Collins, Michael, and Yoram Singer. 1999. "Unsupervised Models for Named Entity Classifi cation." In *Proceedings of the Conference on Empirical Methods for Natural Langauge Processing and Very Large Corpora*. San Francisco, CA: Morgan Kaufmann.

Moldovan, Dan, Sanda Harabagiu, Marius Pasca, Rada Mihalcea, Richard Goodrum, Roxana Girju, and Vasile Rus. 2000. "The Structure and Performance of an Open-Domain Question-Answering System." In *Proceedings of the 38th Meeting of the Association for Computational Linguistics*, 563–570. Morgan Kaufmann.