Adaptive Natural Language Processing

Ralph Weischedel

BBN Systems and Technologies Corporation
10 Moulton St.
Cambridge, MA 02138

1 Objectives

Current NLP technology is very weak at understanding new words, novel forms, or input containing errors. The objective of this project is a pilot study of several new ideas for the automatic adaptation and improvement of natural language processing (NLP) systems. The effort focuses particularly on automatically inferring the meaning of new words in context and on developing partial interpretations of language that is either fragmentary or beyond the capability of the NLP system to understand. The techniques are being evaluated in a message processing domain, such as automatic data base update based on articles from The Wall Street Journal on corporate takeover bids.

The NLP system will use large annotated corpora, such as those being developed under the DARPA-funded TREEBANK project at the University of Pennsylvania, to adapt by acquiring syntactic and semantic information from the annotated examples. Large knowledge bases of common facts will contribute to adaptability by providing information necessary for semantic analysis and discourse analysis. Statistical language modeling, based on probability estimates derived from the large corpora, will provide a means of ranking alternative interpretations of fragments.

This pilot study is designed to test the feasibility of such a new approach.

2 Summary of Accomplishments

In the three months since this project began, we have run pilot experiments on the effectiveness of probability models for (1) ranking interpretations of sentences, (2) predicting the part of speech of known but ambiguous words, and (3) predicting the part of speech of unknown words. Additionally, we are experimenting with using unification algorithms to infer properties of an unknown word from examples.

- In preliminary experiments, we obtained a reduction in the error rate in selecting the correct interpretation of a sentence by a factor of from two to four, depending on the test material.
- Using supervised training for a tri-tag probabilistic model, we achieved a 3-5% error rate on a test set in picking the correct part of speech.
- In a preliminary experiment, we obtained a reduction in the error rate in predicting part of speech of an *unknown* word by a factor of two, compared to random choice.
- We found that the probabilistic language model is useful in indicating which parses containing an unknown word should be used for inferring the new word's properties.
- We installed and evaluated available software that could contribute to this effort, including the Proteus natural language system from New York University and the SMART information retrieval software from Cornell University.

3 Plans

- Implement and test procedures for ranking partial interpretations when the system cannot fully interpret the input.
- Revise probability models and inference strategies to improve system performance.
- Explore automatic methods for learning semantic information of unknown words.