Adaptive Natural Language Processing

Ralph Weischedel

BBN Systems and Technologies
10 Moulton Street
Cambridge, MA 02138

OBJECTIVES

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 uses large annotated corpora, such as those being developed under the DARPA-funded TREE-BANK project at the University of Pennsylvania, to adapt by acquiring syntactic and semantic information from the annotated examples. 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, running from March, 1990 through March, 1991, is designed to test the feasibility of such a new approach.

RECENT RESULTS

We have run pilot experiments on the effectiveness of probability models for (1) ranking interpretations of sentences, (2) predicting the part of speech of words, (3) predicting the part of speech of unknown words, and (4) classifying text. Additionally, we are experimenting with using unification algorithms to infer properties of an unknown word from examples.

1. We obtained a reduction in error rate in selecting the correct interpretation of a sentence by a factor of two compared to no model. A context-free probability model on supervised training of only 80 sentences was used in the experiments.

2. Using supervised training with a tri-tag probabilistic model, we achieved a 3-5% error rate in picking the correct part of speech on a test set including both known and unknown words. As little as 64,000 words of supervised training data was used; with 1,000,000 words of supervised training, less than a 1% improvement in error rate resulted. Therefore, much less training data than theoretically required proved adequate.

3. In processing unknown words, the best errors rate on predicted part of speech as reported in the literature is only 75%. Using a tri-gram model, we found an error rate of 50%. Adding an estimate of the probability of a word ending, given the part of speech, reduced the error rate to 18%. Adding a probability estimate factoring in the likelihood of capitalization, given a part of speech, reduced the error rate for unknown words to 15%.

4. It is well known that a unification parser can process an unknown word by collecting the assumptions it makes while trying to find an interpretation for a sentence. Adding a context-free probability improved the unification predictions of syntactic and semantic properties of an unknown word, reducing the error rate by a factor of two compared to no model.

5. One set of experiments was classification of MUC articles into a specific subcategory or not. We trained a simple classification algorithm to generate a boolean classification tree for each relevant subcategory, that is, a tree which says whether the article is or is not in the category. Given an article, the different classification trees can be applied to it to determine which relevant categories the article is related to. In the following results, "recalled" is the probability that a message in the class would be classified correctly, and "filtered" is the probability that a message not in the class would be classified correctly.

BOMBING	100% recalled, 83% filtered
MURDER	87% recalled, 53% filtered
KIDNAP	76% recalled, 93% filtered
ARSON	97% recalled, 97% filtered

PLANS FOR THE COMING YEAR

- Document results in the project final report.