ROBUSTNESS, PORTABILITY, AND SCALABILITY OF NATURAL LANGUAGE SYSTEMS

Ralph Weischedel

BBN Systems and Technologies
70 Fawcett Street
Cambridge, MA 02138

1. OBJECTIVE

In the DoD, every unit, from the smallest to the largest, communicates through messages. Messages are fundamental in command and control, intelligence analysis, and in planning and replanning. Our objective is to create algorithms that will

1) robustly process open source text, identifying relevant messages, and updating a data base based on the relevant messages;

2) reduce the effort required in porting natural language (NL) message processing software to a new domain from months to weeks; and

3) be scalable to broad domains with vocabularies of tens of thousands of words.

2. APPROACH

Our approach is to apply probabilistic language models and training over large corpora in all phases of natural language processing. This new approach will enable systems to adapt to both new task domains and linguistic expressions not seen before by semi-automatically acquiring 1) a domain model, 2) facts required for semantic processing, 3) grammar rules, 4) information about new words, 5) probability models on frequency of occurrence, and 6) rules for mapping from representation to application structure.

For instance, a statistical model of categories of words will enable systems to predict the most likely category of a word never encountered by the system before and to focus on its most likely interpretation in context, rather than skipping the word or considering all possible interpretations. Markov modelling techniques will be used for this problem.

In an analogous way, statistical models of language will be developed and applied at the level of syntax (form), at the level of semantics (content), and at the contextual level (meaning and impact).

3. RECENT RESULTS

- Consistently achieved high performance in Government-sponsored evaluations (e.g., MUC-3, MUC-4, etc.) of data extraction systems with significantly less human effort to port the PLUM system to each domain, compared with the effort reported in porting other high-performing systems.

- Sped up the PLUM data extraction system by a factor of three.

- Ported PLUM to a microelectronics domain with only seven person weeks of effort. (Typically, systems are ported to a new domain in half a person year or more.)

- Developed a probabilistic model of answer correctness which requires only a set of articles and correct output (the data that should be extracted for each article) as training. This can be used as a model of confidence or certainty on each data item extracted by the system from text.

- Successfully applied a statistical text classification algorithm in MUC-4. The algorithm is trained automatically from examples of relevant and irrelevant texts. The user can specify the degree of certainty desired.

- Distributed POST, our software for statistically labelling words in text, to several other DARPA contractors (New Mexico State University, New York University, Syracuse University, and the University of Chicago).

4. PLANS FOR THE COMING YEAR

Create a probabilistic model for predicting the most likely (partial) interpretation of an input, whether well-formed, novel, complex, or ill-formed.

Develop procedures for automatically learning template fill rules from examples.

Participate in MUC-5 evaluation in all domains.