ROBUSTNESS, PORTABILITY, AND SCALABILITY OF NATURAL LANGUAGE SYSTEMS

Ralph Weischedel

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

OBJECTIVE

In the DoD, every unit, from the smallest to the largest, communicates through messages. Messages are fundamental in command and control, in 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;

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

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

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 semantic 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).

RECENT RESULTS

Achieved performance levels in MUC-3 of identification of over 40% of the data present ("recall") with an accuracy above 50% ("precision"). (Only one quarter of the systems in MUC-3 achieved comparable performance; we achieved this performance with half a person-year of effort to move to this domain, much less than the labor invested by the other top performing groups.)

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

Ported our PLUM message processing system to a class of long range air messages in only seven person-weeks.

PLANS FOR THE COMING YEAR

Create automated procedures for the syntactic training of NL systems, both to improve system performance and to reduce human effort in porting the NL system to a new domain.

Create automated procedures for semantic training.

Develop strategies for automatically inferring a domain model from a corpus, a task which is highly labor-intensive in today's technology.

Create a probabilistic model for predicting the most likely (partial) interpretations of a novel form or errorful input, both of which are significant challenges to current technology.