ROBUSTNESS, PORTABILITY AND SCALABILITY OF NATURAL LANGUAGE SYSTEMS

Ralph Weischedel
weischedel@bbn.com

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. Message 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;

2) reduce the effort required in porting 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 enables 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 are used for this problem.

In an analogous way, statistical models of language are being 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 (MUC-3, MUC-4, MUC-5 and TIPSTER evaluations) 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.

- Achieved very consistent, high performance across both English and Japanese and across both domains (joint ventures and microelectronics) in MUC-5 data extraction performance.

- Applied our probabilistic model of answer correctness to improve the performance of our PLUM data extraction system in MUC-5.

- Achieved speedup by a factor of 80 in POST, our Hidden Markov Model for labeling words in text by part of speech. POST is now distributed to many ARPA contractors (Boston Univ., New Mexico State Univ., New York Univ., Paramax, Syracuse Univ.) and other sites (Advanced Decision Systems, Duke Univ., Univ. of Iowa, City Univ. of New York, and Univ. of Toronto).

- Completed grammar learning experiments showing that the error rate of a stochastic parser is a factor of two less than the same parser without a statistical language model.

- Integrated a pattern matching component into our linguistically motivated framework to give semantics to fragmented parses and discontiguous constituents.

- Created new demonstrations of the PLUM data extraction system in processing English texts about microelectronics and Japanese texts about microelectronics.

4. PLANS FOR THE COMING YEAR

Participate in MUC-6 evaluation at both the application level (extracting data from text) and the understanding level (parsing/semantic/discourse level).

Create/revise probabilistic models for

- word sense disambiguation,
- semantic interpretation, and
- co-reference resolution.

Contribute to the definition of an evaluation methodology for glass box semantic evaluation (Semeval).