RESEARCH IN NATURAL LANGUAGE PROCESSING

Ralph Grishman, Principal Investigator

Department of Computer Science
New York University
New York, NY 10003

PROJECT GOALS

Our central research focus is on the automatic acquisition of knowledge about language (both syntactic and semantic) from corpora. We wish to understand how the knowledge so acquired can enhance natural language applications, including document retrieval, information extraction, and machine translation. In addition to experimenting with acquisition procedures, we are continuing to develop the infrastructure needed for these applications (grammars and dictionaries, parsers, grammar evaluation procedures, etc.).

The work on information retrieval and supporting technologies (in particular, robust, fast parsing), directed by Tomek Strzalkowski, is described in a separate page in this section.

RECENT ACCOMPLISHMENTS

- Developed techniques for computing word similarities based on the co-occurrence of words in the same (syntactic) contexts in a large corpus. Used these similarities to "smooth" automatically-acquired frequency data on verb-argument and head-modifier co-occurrence, and demonstrated that the smoothing increases coverage of the patterns found in new texts. (This work is described in a paper in this volume.)

- Participated in Message Understanding Conference - 4. Incorporated an enhanced time analysis module, an enhanced reference resolution module, and a stochastic part-of-speech tagger into our information extraction component, as well as making general improvements to the semantic models of descriptions of terrorist incidents. Demonstrated a significant improvement in performance over MUC-3.

- In order to gain a better understanding of the problems involved in porting natural language systems to new domains, "translated" our MUC-3/MUC-4 system for extracting information about terrorist incidents to process Spanish news reports. This required development of a relatively broad-coverage Spanish grammar and adaptation of the Collins Spanish-English machine-readable dictionary.

- Developed a prototype procedure for acquiring transfer rules from bilingual corpora through automatic alignment of parse trees in the source and target languages.

- Developed specifications for a common, broad-coverage syntactic dictionary of English (COMLEX).

- Continued participation in a group to define common metrics for grammar evaluation. Applied these metrics to the output of two different NYU parsers (the Proteus parser and the Tagged Text Parser) analyzing a Wall Street Journal corpus.

PLANS FOR THE COMING YEAR

- Participate in Message Understanding Conference - 5. Apply procedures for semantic pattern acquisition from corpora to speed the acquisition and broaden the coverage of the patterns for the "joint-venture" domain.

- Continue work on semantic pattern acquisition procedures. Experiment with larger corpora, with alternative measures of word similarity, and with clustering procedures to identify semantic classes.