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, 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, as well as a paper in this volume.

RECENT ACCOMPLISHMENTS

- Extended earlier work on the acquisition of semantic patterns from syntactically-analyzed corpora, and on the generalization of these patterns using word similarity measures obtained from the corpora. Measured the coverage of the collected patterns as a function of corpus size, and compared this with an analytic model for such coverage.

- Participated in Message Understanding Conference - 5. Substantially extended our lexical preprocessor to identify company names, people's names, locations, etc. Added an acquisition tool for lexico-semantic models, which allows users to specify correspondences between lexical and semantic structures through example sentences.

- Organized meeting for planning of Message Understanding Conference - 6. Coordinated efforts for developing the different corpus annotations which will be required. (These plans and annotations are described in a separate paper in this volume.)

- Developed improved procedures for the alignment of syntactic structures in sentences drawn from parallel bilingual corpora. The goal of this effort is to automatically learn transfer rules for a machine translation system from a bilingual corpus; the starting point is an (incomplete) set of word correspondences from a bilingual dictionary. Demonstrated (using a small Spanish-English corpus) that an iterative algorithm, which uses initial alignments to obtain additional correspondences between words and between grammatical roles, can yield better final alignments. (This work is also supported by the National Science Foundation.)

- Continued studies of appropriate feature structures for a common, broad-coverage syntactic dictionary of English (COMLEX). This work complemented the ongoing effort for creation of COMLEX, which is being supported by ARPA through the Linguistic Data Consortium. (The work on COMLEX is described in a separate paper in this volume.)

PLANS FOR THE COMING YEAR

- Extend earlier work on stocastic grammars for parsing: experiment with alternative word contexts for use in computing conditional probabilities; experiment with alternative search algorithms to obtain speed/precision trade-offs.

- Continue work on semantic pattern acquisition procedures. Experiment with alternative measures of word similarity for use in generalizing patterns extracted from corpora.

- Continue planning for MUC-6. Coordinate efforts to develop specifications and annotated corpora for named entities, predicate-argument structure, coreference, and word sense information; to develop scoring rules for the different evaluations; and to define tasks for MUC-6 dry run in Fall 1994.

- Apply bilingual alignment algorithm to larger corpora. Develop generalization algorithms for transfer rules extracted from bilingual corpus.