RESEARCH IN NATURAL LANGUAGE PROCESSING

Ralph Grishman, Principal Investigator

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

PROJECT GOALS

Our general objective has been the enhancement of techniques for extracting information and retrieving documents from natural language text. Our focus has been on methods which automatically learn syntactic and semantic properties of the language used in particular domains, and on techniques for enhancing the robustness of natural language analyzers. This work covers a number of areas; we summarize our accomplishments and plans for each area below.

INFORMATION EXTRACTION

A major portion of our effort over the past year was devoted to participation in the Third Message Understanding Conference (MUC-3). As the messages involved (newspaper reports of terrorist activity) were significantly more complex than those for MUC-2, a number of enhancements were required to our information extraction system. These included extensions to our grammar, use of a commercial machine-readable dictionary (the Oxford Advanced Learner's Dictionary) as our primary source of lexical information, and several additional techniques for recovery in the event that an entire sentence cannot be analyzed syntactically or semantically. The enhancements were described in detail in the MUC-3 Proceedings. Substantial effort was also required to develop a semantic model of the terrorist domain. The performance of the resulting system compared favorably to others participating in MUC-3. We are currently developing further system enhancements for participation in MUC-4 in June of 1992.

PARSER EVALUATION

To assess alternative techniques for improving parsing performance, we have applied a metric of parsing quality suggested by Ezra Black. Using this metric, we compared parses produced by the Univ. of Pennsylvania Tree Bank for a portion of the MUC-3 corpus against those produced by our system (with some automatic restructuring of our parses to improve their alignment with the Tree Bank). We evaluated a number of methods for improving parser performance, including fitted parses, closest attachment of modifiers, hypothesis merging, stochastic grammars, and stochastic part of speech taggers. Most of these results will be reported at the Third Conf. on Applied Natural Language Processing.

ACQUISITION OF SELECTIONAL PATTERNS

We performed a syntactic analysis of the MUC-3 corpus (automatically, without selectional constraints), and performed a frequency analysis of the co-occurrence patterns to identify the common semantic patterns. When these patterns were used as the basis for selectional constraints in further parsing, they were found to do slightly better than manually prepared constraints. We intend to extend this technique to substantially larger corpora in the coming year.

DOCUMENT RETRIEVAL

We have continued our research on using syntactic analysis to enhance keyword-based document retrieval, both by identifying larger patterns than single words and by automatically discovering term similarity relations from word co-occurrence patterns in a large corpus. When applied to a small corpus of computer science abstracts, modest improvements in performance were demonstrated; these are reported in a separate paper in these proceedings. Over the coming year we intend to apply this technique to much larger corpora as part of the Text Retrieval Evaluation Conference.

MULTI-LINGUAL SYSTEMS

We are continuing our work, sponsored jointly with the National Science Foundation, on Japanese-English sublanguage-based machine translation. We completed a small system for translating programming language texts, which also incorporated the reversible grammar technology we had previously developed. Over the next year we intend to perform initial experiments for discovering transfer rules from parallel bilingual corpora. We are also developing a Spanish version of our information extraction system, in order to better understand the problems of porting such a system across languages.