New York University Proteus Project: ROBUST AND PORTABLE TEXT PROCESSING

Ralph Grishman, Principal Investigator

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

PROJECT GOALS

Our primary goal is the development of robust and portable systems for processing natural language text, particularly for the purposes of extracting information or retrieving passages or documents from a text collection. A major focus has been on automatically or semi-automatically acquiring the syntactic and semantic characteristics of new domains from samples of text.

RECENT WORK

Over the last few months we have begun to adapt our Proteus information extraction system to process reports of terrorist activity. This is being done as part of MUC (Message Understanding Conference) -3, a comparative evaluation of information extraction systems organized by the Naval Ocean Systems Center. These reports are substantially more complex, both syntactically and semantically, than the Navy operational reports we previously processed. Considerable effort has therefore gone into extending our grammatical coverage and improving the efficiency and accuracy of our parser; we have experimented with several techniques, including closest attachment heuristics, merging of alternative analyses of constituents, statistically-trained stochastic grammars, and stochastic part-of-speech tagging prior to parsing (the last done in collaboration with BBN Systems and Technologies Corp.).

We have also begun to investigate the benefits of natural language processing for document retrieval. We have developed a fast and robust Tagged Text Parser, which uses text stochastically tagged by part-of-speech (again with the assistance of BBN), and generates full syntactic analyses (possibly skipping some unanalyzable constituents). These parses are then used to identify co-occurrence patterns and compute similarity coefficients between words. This work, by Strzalkowski and Vauthey, is reported in a separate paper in these proceedings.

Finally, as part of our research on sublanguage-based machine translation, we are continuing development of our Japanese grammar.

PLANS FOR THE COMING YEAR

Our primary task for the remainder of MUC-3 (through May 1991) will be to incorporate additional semantic information about the terrorist domain. We intend to do this, as much as possible, through semi-automatic techniques driven by the parsed corpus of 1300 reports. Following completion of the formal evaluation in May, we expect to spend considerable time evaluating the contributions of various system features to our overall performance.

For document retrieval, we intend to utilize a clustering procedure which, based on the similarity coefficients we have generated, will form domain-specific word classes. We will then investigate the benefits of using these word classes as a thesaurus for keyword-based document retrieval, using a standard document test collection.

Finally, we expect over the coming year to bring together our pilot study of sublanguage-based machine translation, our work on reversible grammars, and our Japanese and English grammars to create a reversible Japanese-English translation system, albeit initially operating on a very limited domain and vocabulary.