ROBUST NATURAL LANGUAGE ANALYSIS

New York University

Principal investigator: Ralph Grishman

Our basic goal is the development of more robust systems for extracting information from natural language text. A robust system is one which is able to extract at least partial information despite the presence of ill-formed or unexpected syntactic, semantic, or discourse structures. Our approach has two aspects: First, we incorporate a rich set of syntactic, semantic, and discourse constraints, so that one type of constraint can guide us to a correct analysis even if another type is violated. Second, we provide mechanisms for relaxing individual constraints, and for scoring alternative analyses, so that the analysis violating the fewest constraints (and therefore, presumably, the best analysis obtainable for a sentence) will be selected.

Our vehicle for testing this approach is a system for analyzing Navy operational messages. It was originally developed for RAINFORMs and last spring was adapted to process OPREPs as well; both are messages describing naval sightings and engagements. We have continued developing a grammar for analyzing these highly telegraphic messages. This grammar relies on weighted syntax rules, which allow for a wide variety of omissions but prefer analyses as full English sentences (this approach was described at the February '89 DARPA conference). We have developed a semantic classification hierarchy and set of lexico-semantic models which characterize the types of arguments and modifiers which can occur in clauses and noun phrases in these messages. We have also developed a simple discourse model, which identifies plausible event sequences within an analyzed message.

The weighted syntax rules (which allow for a number of variations from standard syntax) and the discourse rules (which prefer but do not require particular event sequences) have increased our system robustness. We have recently incorporated a number of additional techniques to increase robustness. These include a spelling corrector and a prefix parse mechanism (which, if no analysis can be obtained of the entire sentence, will take the longest substring of the input, beginning with the first word, for which an analysis has been obtained). We have arranged the lexico-semantic models in a hierarchy, so that if the semantic model for a specific word does not match, an attempt will be made to match a more general model. And perhaps most importantly, we have begun to use preference semantics: if no analysis of the input satisfies all semantic constraints (all clauses and phrases match some semantic model), the analyzer will seek an analysis violating the fewest constraints. This approach is described in an accompanying paper in this volume.

The system incorportaing these robustness measures was evaluated as part of the MUCK-II Message Understanding Conference, held at the Naval Ocean Systems Center (San Diego) in June 1989. For each message narrative, the systems under evaluation had to identify the types of events (detection, attack, etc.) and the critical parameters of the event (agent, object, time, location, etc.). We believe that the robustness techniques just described were crucial to whatever modest success we achieved with the MUCK-II data.

Over the coming months we intend to analyze the individual contribution to robustness made by the various techniques mentioned above, as a step towards gradually refining these techniques. In addition, we want to move from constraints and weights which are set by hand to ones obtained automatically from a sample text corpus. We have already conducted some preliminary experiments, with encouraging results, for obtaining the weights for a (stochastic) context-free grammar for message analysis from data on the frequency with which the productions are used in a sample corpus. We also intend to continue some earlier work, using larger text samples, on the acquisition of selectional constraints from parsed text samples.