Research in Text Processing: Creating Robust and Portable Systems

Computer Science Department
New York University

Ralph Grishman, Principal Investigator

Objective

Our goal is to improve the technology for retrieving passages, extracting specific facts, and creating formatted data bases from large text collections. In particular, we are concerned with developing techniques for automatically training language processing systems to the syntax and semantics of particular domains and types of text in order to improve system performance.

Approach

Improving the natural language technology for information extraction and retrieval will require

- increased knowledge of specialized language usage and domain semantics
- tools for acquiring this knowledge
- analysis mechanisms which can cope with gaps in this knowledge

In natural language text, much of the information is implicit and much of it, viewed in isolation, is ambiguous. Increased information about syntactic usage, discourse patterns, and the semantics of particular domains is essential to resolve this ambiguity and extract the intended facts from the text.

However, collecting this information manually for each type of text is difficult and time-consuming, and renders the system non-portable. It is therefore desirable to be able to extract such characteristics as the relative preference for different syntactic structures and the semantic classes and constraints automatically from a sample of text in a particular domain.

Since the text samples are finite, this information will always be incomplete. In addition, any real text will contain typographical and syntactic errors and semantic relations outside the principal domain. In consequence, a high-performance system will require a forgiving analysis procedure which tries to minimize constraint violations but does not insist on a "perfect" input.

To guide and evaluate our work on the underlying technologies, we have developed three message processing applications over the past five years. The first was for CASREPs – equipment failure messages. The focus for this system was on deep domain models for language understanding, and in particular for the determination of the implicit causal and temporal relations between events in a narrative. The other systems involved RAINFORMs and OPREPs – messages describing naval encounters and engagements. These systems were developed for the Message Understanding Conferences organized by the Naval Ocean Systems Center. The focus for these systems was on robustness: the ability to extract at least partial information despite violations of syntactic or semantic constraints.

Recent Work

Since the last Message Understanding Conference (in June 1989) we have analyzed some of the factors contributing to the performance of our system. We reported at the last DARPA Workshop on the crucial role played by preference semantics in increasing the number of events correctly identified. We have also examined the effects of our heuristics on the error rate – the number of incorrect event frames generated by the system. We found that the number of errors increased only slightly in absolute terms, and decreased as a fraction of the total set of event frames generated by the system (Grishman and Sterling, COLING 90 Proc.). We have also improved system speed somewhat, reducing the time required to process the 105 message corpus from about 14 hours to less than 4 hours (on a Symbolics 3645).

As part of our progression towards more knowledgeable systems, we have developed some techniques for incorporating information about the relative frequency of different syntactic constructs into the context-free core of our grammar. This work is being reported at this workshop (Chitrao and Grishman).

We have incorporated a commercial dictionary (the Oxford Advanced Learner's Dictionary of Contermporary English) into our system to supplement the specialized dictionaries we have developed for particular applications.

Finally, we have continued our work on sublanguage-based Japanese-English machine translation (supported in part by the National Science Foundation).