Active Knowledge Structures in Natural Language Understanding

Computing Research Laboratory
New Mexico State University
Principal Investigator: Yorick Wilks
yorick@nmsu.edu
(505) 646-1835

PROJECT GOALS

To investigate a theory of message and discourse understanding based on the building of explanatory causal models and the nested beliefs of discourse agents, so as to construct robust systems for message understanding with wider application to machine translation and text retrieval.

APPROACH

We view the task of the robust understanding of messages from texts and discourse as the use of extraction of gists from a noisy background by using techniques of (a) the recursive computation of agent's points of view of each others' environments, beliefs, expertise etc. and (b) representations of those beliefs and expertise as networks which are obtained by "best-fit" methods against stored knowledge structures. We implement these high-level assumptions by means of different types of parsers (some more syntactic, others more semantic) and the choice between these we see as purely empirical, based on our evaluation methodology. At present we are concentrating on PREMO-II, a semantics based parser which does contain some syntactic rules, but which can be weighted based on statistical surveys of Navy message texts.

RECENT WORK

We have continued o work within the Navy message domain with this project since we were late starters and wanted to get some results before switching over to new domains and text types (see below). We have succeeded in integrating the ViewGen belief manipulation system with the conceptual graph + MGR knowledge representation, so as to provide a single system that can both guide the message parsers and represent the results of message extraction.

This rich representation has, in turn, been linked to the PREMO-II semantics-based parser that parses Navy messages on the bases of preferences, anticipated metonymies closely linked to preferences (e.g.CONTACT WENT SINKER), and the weighting of simple syntactic rules based on statistics of a large sample of such messages. This type of method, although derived for general text has proved very suitable for "systematically ungrammatical" text like the Navy messages.

We have begun to investigate how to expand these techniques to a different text type, the longer terrorist messages, in conjunction with the set of techniques we have proposed for the TIPSTER extraction program.

Some of our effort under this contract has been diverted to getting the ACL Consortium for Lexical Research up and running before its official DARPA funding date of March 1st.

PLANS FOR THE COMING YEAR

We expect to take the shell of the PREMO-II method and provide it with a new data-base of lexical entries, partly automatically derived from existing machine readable dictionaries and partly tuned against large text data-bases. Our standards of content in the lexical entries for parsing will be those of the Pustejovsky and Annick work at Brandeis University, with whom we expect to collaborate. The method will also be integrated with surface demons for categories like proper names, country and place names, company names, etc. that have been derived separately by David MacDonald. These methods will give, we believe, a general and robust method of extraction from documents on a large scale.