EXTENDING THE SCOPE OF TEXT UNDERSTANDING SYSTEMS EVALUATION

Beth Sundheim
Naval Ocean Systems Center, Code 444
San Diego, CA 92152-5000
sundheim@nosc.mil

The Naval Ocean Systems Center is extending the scope of previous efforts in the area of evaluating English text analysis systems. These evaluations are intended to advance our understanding of the merits of current text analysis techniques, as applied to the performance of a realistic information extraction task (where the output resembles the contents of a formatted database). The premise on which the evaluations are based is that task-oriented tests enable straightforward comparisons among systems and provide useful quantitative data on the state of the art in text understanding. Furthermore, the data can be interpreted in light of information known about each system's text analysis techniques in order to yield qualitative insights into the relative validity of those techniques as applied to the general problem of information extraction.

The third evaluation is being launched in August 1990. It will be broader in scope than previous ones in most respects, including text characteristics, task specifications, and anticipated range of analysis techniques. These differences are motivated by feedback from participants in the second evaluation. Two thousand texts are being collected for use as training and test data; a data extraction task on the subject of Latin American terrorism and a semi-automated scoring algorithm are being defined.

The conference announcement has been sent to system developers who are pursuing a wide range of text interpretation techniques (e.g., statistical, key-word, template-driven, and natural language processing), and it is hoped that at least twenty organizations will participate. A conference will be held in mid-November, 1990 to discuss the results of a trial test run and to work out issues affecting the test design, scoring, etc. A shorter conference will be held in mid-February, 1991 to present the results of final testing and summarize the findings of the evaluation.

All systems will be evaluated on performance on the database generation task in a blind test. Measures will be database completeness (recall) and accuracy (precision), which will be calculated for the test set overall, with breakdowns by slot. It is expected that different techniques will have varying degrees of success in filling slots, depending on such factors as whether the number of possible slot fillers is small, finite, or open-ended and whether the slot can typically be filled by fairly straightforward extraction or not. In addition to these official measures, unofficial measures will be taken of performance on particular linguistic phenomena (e.g., conjunction), as measured by the database fills generated by the systems in particular sets of instances.