EVALUATING TEXT UNDERSTANDING SYSTEMS

Beth M. Sundheim

Naval Command, Control and Ocean Surveillance Center
(formerly Naval Ocean Systems Center)
Naval Research, Development, Test and Evaluation Division, Code 444
San Diego, CA 92152-5000

PROJECT GOALS

The objectives of this project are to advance our understanding of the merits of current text analysis techniques, as applied to the performance of realistic text analysis tasks, and to achieve this understanding by means of a sound performance evaluation methodology. The performance data can be interpreted in light of information known about the text interpretation techniques for the various systems to yield qualitative insight into the relative validity of those techniques for the text analysis task. The data can also be used as a means for determining which research areas are most critical to the successful performance of the task.

The most recent performance evaluation was conducted in May, 1991, on systems contributed by 15 R&D sites. The evaluation task was intended to yield insight into text analysis technology, including the use of information retrieval technology (document retrieval and categorization) instead of or in concert with language understanding technology. The evaluation concluded with the Third Message Understanding Conference (MUC-3).

MUC-3 RESULTS

* Significant performance benchmarks that show substantial capability for the top-scoring systems, given the extreme difficulty of the task: the systems demonstrated an ability to extract up to approximately 40-50% of the information expected and to extract information with at least 50-60% accuracy.

* A rich set of data on each of the systems for further analysis, a highly improved set of performance metrics embedded in a flexible, semiautomated scoring program, and a large database of texts and extracted information that can be used to support future research in computational linguistics.

* Useful insights into the merits of the various technologies, including these high-level ones: (1) although the top-scoring systems reflected a diversity of overall approaches, they all employed robust parsing methods and domain-specific knowledge; (2) systems using only nonlinguistic techniques were not able to score as well as those that included linguistic techniques.

* Identification of the need for scientific breakthroughs in at least one major research area, namely discourse (tracking the flow of an incident description across sentences and paragraphs).

CURRENT EFFORTS

* Modify the evaluation task and testing techniques to improve the reliability and utility of the results.

* Explore new ways of using the existing framework to gain insight into particular aspects of performance.

* Measure progress in the field by running the evaluation again in June, 1992, using a new test set, an improved scoring program, the refined basic evaluation task, and any newly-defined subtasks.

* Use the evaluation methodology as the basis for developing an evaluation plan for the information extraction portion of the DARPA Tipster natural language program.