Data Collection and Evaluation II

Ralph Grishman
New York University

This session consisted of four presentations concerning data collection and the evaluation of speech and text processing systems.

The first presentation, by Dave Pallett, described an evaluation by Pallett, Fiscus, and Garofolo (National Institute for Standards and Technology) of the significance of differences in performance of various speech understanding systems. A number of sites which are developing such systems have been using the DARPA Resource Management Speech Corpora to evaluate their systems. As performance gradually improves, and differences — between systems and between successive versions of a system — narrow, the question arises as to whether reported differences are statistically significant. To assess this question, the authors have implemented two statistical tests and applied them to the output of several speech understanding systems. They reported that, for speaker dependent systems and systems using a word-pair grammar, differences were often not significant; for speaker independent systems and systems not using a grammar, significant differences were found in many cases.

Turing to issues of data collection, Mitch Marcus presented a report by Marcus, Santorini, and Magerman (Univ. of Pennsylvania) of "First Steps Towards an Annotated Database of American English." This is an effort to collect and annotate, over several years, substantial samples of written and spoken English. Work in the first year (just beginning) will be on written text. Marcus described the development of guidelines and tools for the text annotation. A small set of parts of speech (34) has been defined, an annotation workbench based on the Gnu Emacs editor has been developed, and a manual has been prepared. Experiments with several approaches to part-of-speech tagging indicated that statistically-based automatic tagging (based on work by Ken Church) followed by manual review was the most efficient and reliable. Tagging rates of 20 minutes/1000 words, with error rates of 3-4%, were reported.

Mark Liberman then reported on efforts of the ACL (Association for Computational Linguistics) Data Collection Initiative. This initiative aims to acquire and prepare a large text corpus, to be made available without royalties for scientific research. The text will be formatted using SGML (the Standard Generalized Markup Language). To date several hundred million words of text have been collected, including material from newspapers, parliamentary records, literary works, technical abstracts, and reference works. An initial release of a subset of this material is planned for the near future.

To close the session, Beth Sundheim of the Naval Ocean Systems Center reported on MUCK-II, the second Messsage Understanding Conference, held at NOSC (San Diego) in June 1989. This conference was the first substantial effort at measuring and comparing the performance of message understanding systems. The task given to these systems was to identify — from brief narratives of naval encounters — certain types of critical events, and to extract the agent, instrument, time, location, etc. of each event. A total of 130 messages were prepared for training and evaluation, along with specifications of the correct output for each message. Nine groups participated in the evaluation procedure. The results from the training and on-site evaluation test sets were briefly presented at the session; more detailed results will be included in a forthcoming NOSC report.

The session highlighted two trends among the work in computational linguistics. First, the gradual maturing of evaluation efforts — in particular, the shift in the text processing domain from 'proof of concept' to quantifiable evaluation. Second, the increasing interest (noted by Mitch Marcus at several points in this conference) in the use of much larger data bases (of text, speech, and lexical information) as basic tools in computational linguistics research.