SESSION 2: INVITED OVERVIEWS

Madeleine Bates

BBN Systems & Technologies
70 Fawcett Street
Cambridge, MA 02138

By adopting the name Human Language Technology for one of its flagship programs, ARPA intended to included everything that is involved in understanding and/or generating natural human language. In the Information Age, more and more resources are being devoted to collecting, retrieving, and using information (text and speech) in digital form. The time is ripe to multiply the power of human-machine problem solving systems by automating ways to cope with the information explosion we are all experiencing.

The scope of this workshop, expanded from a previous concentration on spoken language systems to include all spoken and written language work at ARPA, meant that many of the attendees were coming into contact for the first time with researchers from related but different disciplines. Several invited overview presentations helped to set the context for what was to come later in the workshop.

Beth Sundheim of the Naval Command, Control & Ocean Surveillance Center described the series of four MUCs (Message Understanding Conferences) that have taken place since 1987. These conferences, each involving a highly structured evaluation of message processing systems, have served to quantify the state of the art in this field, and to provide a forum for studying and modifying the evaluation methodology. Participation in the MUCs has risen from 6 sites to 17, demonstrating increased interest in message processing as a task, and increased willingness to evaluate systems using a common task for which training data is available.

Donna Harman of the National Institute of Science and Technology (NIST) presented the background leading up to the first Text REtrieval Conference (TREC-1), and the results of that conference. The TREC participants were eager to test a variety of retrieval techniques in a common, challenging evaluation. Approaches ranging from pattern matching to term weighting to natural language processing competed in what was probably the best modern information retrieval test, and certainly the largest. Papers by several of the TREC participants appear in this proceedings (particularly in Session 12, Information Retrieval).

Thomas Crystal from ARPA presented an overview of the TIPSTER program, including both the message detection and data extraction tasks. By "detection" is meant two variants of information retrieval: retrospective retrieval and routing. The detection problem is to process queries in the form of user-need statements and finding messages that meet the needs (e.g, are on a particular topic) from among a huge set of messages, some of which are similar but not on the desired topic, and some of which are completely irrelevant. By "extraction" is meant extracting specific types of information from messages that are likely to be relevant to the particular topic. Extraction systems typically apply text understanding techniques to process the text, and then produce database fill from the results of that understanding.

Part of the challenge of the TIPSTER program has been to provide a harder problem than has been worked on before, one that can be solved only by applying very advanced technology; to this end, TIPSTER detection and extraction work has been pursued in both English and Japanese. The current TIPSTER contractors' work is represented throughout this proceedings. Because Tom Crystal will be leaving ARPA shortly, questions about this program should be directed to George Doddington.

George Doddington of ARPA provided an overview of the Spoken Language Systems (SLS) program, which is also well-represented in this proceedings, particularly in Sessions 1 and 3. The SLS program is concerned with all aspects of human-machine communication by voice, and has been using ATIS (Air Travel Information System) as a common domain for development and evalulation. The paper by David Pallett gives a summary of the most recent benchmark evaluations in this program.

Inquires about the SLS program or any other aspect of the ARPA HLT program should be addressed to:

Dr. George Doddington
ARPA / SISTO Room 744
3701 N. Fairfax Drive
Arlington, VA 22303-1714
gdoddington@darpa.mil
703-696-2259