Session 5: Overview of the ATIS System

David S. Pallett, Chair
National Institute of Standards and Technology
Bldg. 225, Rm A216
Gaithersburg, MD 20899

Two Sessions were devoted to introducing the DARPA SLS Common Task: the Air Travel Information System (ATIS). The first of these (in the late afternoon) provided an overview of the ATIS task and presented a summary of results reported to NIST in the first of a projected series of ATIS benchmark tests, while the second (immediately after dinner) served to describe different approaches in implementing the common task at several sites: BBN, CMU, MIT/LCS, SRI and Unisys. The two sessions were filled with a sense of excitement and appreciation of the precedent-setting nature of these tests, and revealed strong differences of opinion over details of implementation of the common task and implementation of the test, as well as frustration with the limitations on available time and the amount of ATIS domain material at the present time.

The Chairperson noted that in some sense the DARPA SLS community has recently been involved in a "construction project" — building a common task domain that is to rely on a common relational (knowledge) database for database queries, building a speech corpus that is to be used for system development and evaluation, and (perhaps most significantly in the context of these sessions) building an evaluation methodology for DARPA spoken language systems. Such a construction project is not without pressures of scheduling, various constraints and unknown dangers!

The first paper, presented by Patti Price of SRI, outlined some of the issues and rationale behind selection of the air travel domain for a DARPA SLS common task. One important function of the common task is to serve to limit the domain-specific-sub-language for the DARPA spoken language systems. Patti noted that choice of air travel planning makes use of data derived from the on-line Official Airline Guide put into a relational format. It offers the advantages of having been used by "hundreds of thousands of people, providing a wide pool of users familiar with the domain, and the domain is rich, interesting, and can be scaled with the technology" [1].

Charles Hemphill, the "Wizard" at TI responsible for collection of the ATIS Pilot Corpus, next described the process of corpus collection at TI. Spontaneous speech was collected using a Wizard-of-Oz" simulation. Following collection of the speech, TI developed transcriptions, classified the queries, generated reference SQL expressions, and obtained reference answers, prior to sending the corpus to NIST for distribution to the SLS community. Charles noted that our collective experience with the ATIS Pilot Corpus has demonstrated that "objective evaluation of spoken language systems is both possible and beneficial" [2].

Lyn Bates, of BBN, presented a paper describing the development of the evaluation methodology proposed for spoken language systems and implemented for this meeting [3]. Boisen *et al.* had outlined such a procedure at the October 1989 DARPA Workshop [4], and Lyn outlined some of the details of the present implementation for the ATIS domain. Strengths of the methodology noted include: (1) forcing agreement on the meaning of critical terms, (2) being objective [with some caveats], (3) being [largely] automated, and (4) extendibility to other domains and to account for context-dependency and limited dialogue. Weaknesses that have been noted include: (1) not distinguishing between "acceptable" answers and "very good answers", (2) potentially crediting systems that "over-answer" for answers to specific questions (by virtue of providing what some have termed "the kitchen sink" in response to some questions — in the hope that the correct answer will be found somewhere in the response), and (3) not being able to tell if the right answer was gotten for the wrong reason.

Lynnette Hirschman outlined a proposal for automatic evaluation of discourse [5]. The proposal, termed "Beyond Class A", suggests procedures that might be used to permit evaluation of context-dependent queries. To extend the present scoring methodology to evaluate context-dependent sentences, a "canonical display format" could be defined, with reference to this display as an additional input for "resynchronization" of the SLS systems by providing "the full context".

Prior to this meeting, NIST had distributed a set of 93 test ATIS domain queries and received results (more-or-less in "canonical answer format") for a total of 7 systems from 5 sites. These results had been scored at NIST. In the last presentation before dinner, results were distributed by NIST for the preliminary scoring of these results [6]. Bill Fisher reviewed NIST's experience in administering these tests. Bill noted some inconsistencies in different sites' implementations of the tests, and outlined steps NIST had taken to make the scoring software more tolerant of format-related errors. Several sites had opined that three of the test queries were ambiguous, and NIST agreed to delete these 3 from the original test set of 93, leaving a set of 90 "official test queries".

Following the NIST presentation, a break for dinner took place.

REFERENCES

[1] Price, P. , *Evaluation of Spoken Language Systems: the ATIS Domain* (in this Proceedings).

[2] Hemphill, C. T., Godfrey, J. J. and Doddington, G. R., *The ATIS Spoken Language Systems Pilot Corpus* (in this Proceedings).

[3] Bates, M. and Boisen, S., *Developing an Evaluation Methodology for Spoken Language Systems* (in this Proceedings).

[4] Boisen, S. *et al.*, *A Proposal for SLS Evaluation* in Proceedings of the Second DARPA Speech and Natural Language Workshop (Cape Cod, MA) October, 1989.

[5] Hirschman, L. *et al.*, *Beyond Class A: A Proposal for Automatic Evaluation of Discourse* (in this Proceedings).

[6] Pallett, D. S. *et al.*, *DARPA ATIS Test Results: June 1990* (in this Proceedings).