Session 6: ATIS Site Reports and General Discussion

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

Many of the evening session papers served to describe local implementations of ATIS domain systems and to provide local "glass-box" diagnostic evaluations to complement the NIST "black box" scoring. In some cases, these had the appearance of "nit-picking". However, since our collective intent was to develop and implement an evaluation procedure, the criticisms were generally constructive, and each site has much to contribute toward improved implementations of SLS performance evaluation procedures. Many of these observations underscore the complexity of developing standard test procedures in a new and challenging arena: spoken language systems.

In the first paper in the evening session, Sean Bates reported on the BBN ATIS System [1]. Results were cited for two different systems, one of which was the commercial "Parlance" system. The other, the "Delphi" system, is intended to comprise the research NL component of BBN's HARC (Hear And Recognize Continuous speech) system. BBN's diagnostic evaluation indicated that the production of incorrect answers was not a significant problem for the Delphi System. However, there were 38 (out of the 90 "official" queries in the test set) for which the NO_ANSWER response had been given. BBN's analysis outlined the major causes of these NO_ANSWER responses: (1) word senses not previously encountered and (2) lack of inference. Boisen concluded by noting that: (1) "This evaluation methodology works!", (2) more training data and more time to use it is needed, (3) more careful "definition of answer criteria is needed", and (4) that (particularly in view of the low intra-speaker variability in linguistic structure noted in the ATIS Pilot Corpus) the ATIS language is not "varied enough", so that it "would increase the validity of tests of SLS systems if more than one domain were used".

The CMU system to understand spontaneous speech has been termed "Phoenix", and was introduced by Wayne Ward, in the second evening presentation [2]. For text input, a frame-based parser is used to process ill-formed text. In Ward's "diagnostic evaluation", he noted that "a significant number of utterances that parsed were scored as incorrect. Most of these were of two types that resulted from a misunderstanding on [Ward's] part as to what was to be generated." These involved dates and abbreviations. The CMU implementation was unique among those for which results were reported at this meeting in that input speech (waveform) files were processed in addition to the text (SNOR format) files. In this case, output from the Sphinx speech recognition system was passed to the parser, using a word-pair grammar with a cited perplexity of 85.

In questions following Ward's presentation, and in noting the "misunderstanding on [Ward's] part as to what was to be generated" that led to answers "scored as incorrect", Patti Price asked "What is the moral of this story?". To everyone's amusement, particularly those who had been involved in the disputes about the proposed test protocols prior to the test, Ward responded that now "he'll read the net mail!"

Preliminary ATIS development work at MIT was described by Stephanie Seneff [3]. In the MIT ATIS system, "low level functions typically fill slots in an event frame with ... semantic information. Once the entire sentence has been processed and the history frames have been merged, an IDIL query is then constructed from the completely specified query." [In this case, the IDIL query makes use of the Intelligent Database Interface (IDI) "as an intermediary to SQL", provided by researchers at Unisys.] There had been four releases of incremental portions of the ATIS Pilot Corpus, and the MIT group monitored progress in handling the utterances in each successive release, both in terms of parser coverage and agreement of the back-end responses with the canonical answers. These studies led Seneff to express concern "that rules created to deal with utterances in one release don't seem to generalize well to new releases", a finding that may be related to other observations about the "very high inter-speakers variability that accompanies low intra-speaker variability in linguistic structure" (see, for example [1]). While noting that an inordinate amount of time had been required to work with the back end and the need to generate SQL queries, Seneff remarked that "the idea of a common task involving booking flights is a good one", and that they "look forward to ... integrating the natural language component with a recognizer".

In a refreshing contrast to the other papers in the evening presentation (which focussed largely on diagnostic evaluations), Patti Price reported on studies at SRI (involving the ATIS relational database) which assessed the effects of changes in the simulations on the speech and language of the experimental subjects [4]. The stated goal of these studies is to "design an appropriate human-machine interface". Price also noted that "the greatest source of variability in the system is that across subjects". Five experiments were described for several data collection conditions. The SRI studies suggest that "the goal of designing an appropriate spoken language system can conflict with the goal of collecting data for evaluation of spoken database queries", but that they "believe that it is possible to find some ways to coordinate the two endeavors."

In the last of the formal presentations in the session, Lew Norton described the Unisys ATIS domain system [5]. The Unisys approach combines a number of elements: (1) the MIT SUMMIT speech recognition system, (2) the Unisys PUNDIT language understanding system, (3) use of a module termed QTIP (Query Translation and Interface Program, and (4) the Intelligent Database Server (IDS), a "general knowledge/database interface" to mediate access to the database, (5)INGRES to access the ATIS relational database, and (6) a Dialogue Manager to integrate overall user-system interaction. [Note that the MIT system described by Seneff also made use of elements of the IDS component (i.e., the IDI portion).] In the Unisys "diagnostic evaluation" as reported by Norton, errors were noted due to several causes: (1) words not being in the lexicon, (2) problems in parsing, (3) problems in obtaining an appropriate semantic/pragmatic analysis, and (4) failure of the QTIP module to generate an appropriate call to the relational database. Like other sites, the Unisys researchers noted great inter-subject variability — with their systems's "success rate" for the different subjects in the test set ranging from "30% to 88%". Norton further noted that the implications of this finding suggest that there "are a large number of different ways to ask essentially the same questions", and that "a natural language understanding system will have to be trained on much larger volumes of data." This observation was further supported by data documenting the rate of incremental growth of the grammar in the ATIS domain, which appears to be much slower for ATIS than for the MIT Voyager domain.

Following the presentations from BBN, CMU, MIT, SRI, and Unisys, some time was devoted to general discussion of issues raised in the afternoon and evening ATIS Sessions.

Bob Moore underscored what a number of individuals had noted: that there had been insufficient time between the release(s) of the training data and the test data. [It is important to note that the relational ATIS database used in these studies had not been "frozen" until mid-April, and incremental releases of the training data took place during a six-week period during May and just prior to the release of the test data on June 15th.]

Lynnette Hirschman noted that substantially more data should be made available for this domain — of the order of ten times more than to date.

Victor Zue noted that proposals to extend the test methodology to accommodate context (such as those outlined by Bates and Hirschman) seemed attractive, but that all evaluations are, to some degree, subjective and that we need to plan on developing procedures for formal subjective evaluations. Victor likened the present approach to ATIS implementations to a "shotgun" approach, and expressed a preference for more focused or constrained scenarios and local implementations that might be regarded as "rifle" approaches. Victor also underscored what others had suggested: that a pooling of data from several sites may be the only practical way to gather the amount of data that appear to be needed.

John Makhoul noted that the focus of the studies reported at this session should be seen as a vehicle for NL/SLS evaluation, not so much as an effort to develop real air travel information systems.

Patti Price noted "TI's Heroic Role" in developing the ATIS relational database used for these studies, collecting spontaneous speech data and providing "canonical answers". Charles Hemphill and his colleagues at TI worked very hard to provide data for the Pilot Corpus for both training and test purposes, and the ATIS studies at the several sites would not have been possible without the TI group's efforts.

Charles Wayne closed the discussion by thanking the participants for the significant Spoken Language Systems progress in the ATIS task domain.

REFERENCES

[1] Bates, M. *et al.*, *BBN ATIS System Progress Report – June 1990* (in this Proceedings).

[2] Ward, W., *The CMU Air Travel Information Service [sic]: Understanding Spontaneous Speech* (in this Proceedings).

[3] Zue, V. *et al.*, *Preliminary ATIS Development at MIT* (in this Proceedings).

[4] Bly, B. *et al.*, *Designing the Human Machine Interface in the ATIS Domain* (in this Proceedings).

[5] Norton, L. M. *et al.*, *Management and Evaluation of Interactive Dialogue in the Air Travel Domain* (in this Proceedings).