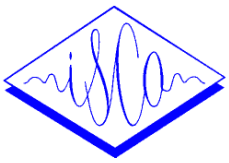

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

THE MIT ATIS SYSTEM: PRELIMINARY DEVELOPMENT, SPONTANEOUS SPEECH DATA COLLECTION, AND PERFORMANCE EVALUATION[1]

Victor Zue, James Glass, David Goodine, Lynette Hirschman,
Hong Leung, Michael Phillips, Joseph Polifroni, and Stephanie Seneff

Spoken Language Systems Group, Laboratory for Computer Science
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA

ABSTRACT

This paper describes the MIT ATIS system, with particular emphasis on the discourse and dialogue models and the consequences for data collection. The ATIS system provides air travel information and can simulate the booking of a flight, using a mixed-initiative dialogue framework. An intermediate semantic frame representation serves as the focal point for all back-end operations, and the discourse model includes the resolution of explicit anaphoric references and indirect and direct references to information mentioned earlier in the conversation. We have collected over 4500 utterances from subjects using the system to solve simulated booking scenarios. We have studied these dialogues, and have used them productively to guide the development of better discourse and dialogue models. We have also tabulated some interesting differences between our data and those collected at Texas Instruments using a very different paradigm.

INTRODUCTION

ATIS, or Air Travel Information Service, is the designated common task of the DARPA Spoken Language Systems (SLS) Program [6]. It is an air travel information system that is designed to provide travel assistance using spoken input. It currently knows about only 11 cities (9 airports) in the U.S., and 8 airlines serving these cities. The system can answer questions about such topics as departure and arrival times of each flight, the type of aircraft used, and meals served. In addition, the MIT version can guide the user through making a flight reservation.

Detailed descriptions of the MIT ATIS system and the discourse/dialogue model have been described previously [7,8]. This paper will focus on the data collection aspects, both on issues of how to do data collection, and on how system improvement can be coupled with the data collection process itself. Our data collection procedure is an involved human/machine interaction, making full use of the existing back-end component. The data that are collected tend to be well matched to the type of data one would expect to see in real system usage.

We will first give a description of the general system architecture, followed by some more detailed discussions of the discourse and dialogue models. We will then document our data collection procedure, and provide some statistical evidence for differences between our data and data collected previously by Texas Instruments (TI) using an "intelligent" wizard paradigm. We will conclude by presenting performance evaluation statistics on two test sets and discussing the use of the data collection process to shadow system development.

[1]This research was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

SYSTEM ARCHITECTURE

The natural language and response generation components of the system make use of a simple semantic frame representation of the meaning which serves as the input for database access, spoken response generation, and history management. The frame design is flexible enough to be readily extended to other domains. Domain-dependent aspects of the system are entered mainly through table-driven mechanisms, with very little explicit programming required.

Processing of a sentence involves several steps. The first step is to provide a parse tree for the input word stream. The parse tree is processed to yield a semantic frame, which is then integrated with available frames from the history. Both an SQL query and a generated text response are derived from the completed frame. The verbal response is spoken to the subject and a table is retrieved from the database through the database management system ORACLE. A table post-processing step converts the table to a much more readable and informative form prior to display. Finally, the system examines the goal plan and optionally initiates an additional response, based on its assessment of a likely next step. More details about this process can be found in [7,8].

Discourse Model We initially designed a relatively simple discourse model that could reasonably handle the majority of situations. The model has since been improved, based on observations from the data, to the point where the present system usually correctly interprets sentences in context.

A very common occurrence within this particular domain is a form of ellipsis in which various modifiers on flights are assumed to carry over from the history without explicit anaphoric reference. Thus, for example, a user typically identifies a source and destination in an introductory sentence, and then assumes that these specifications will carry over to several subsequent sentences. Individual modifiers are inherited unless new modifiers override their inheritance. Exactly which modifiers should block which others was determined empirically from subject data through the data collection episodes. History elements are stored in the standard frame format, and inheritance of a modifier usually amounts to simply inserting it into the appropriate frame of the new sentence.

The newest version of the system retains several distinct flight events in the history. These include the most recent singular set, the most recent plural set, and a pointer to the most recent frame associated which each flight number that has shown

up in a table. This allows users to refer back to previously mentioned flight numbers without specifying their source and destination. The system also retains the most recent *entire clause*, which may be retrieved and modified for clarification questions, as in the sequence, "what meals are served on flight 201?" followed by "How about flight 94?"

When an empty table is retrieved from the database, the verbal response must be modified to reflect possible presupposition failure. For instance, it is inappropriate to say, "No, the flight that leaves at noon does not serve lunch," when in fact there is no flight leaving at noon. Similarly, when queried, "What is the aircraft for flight 94," the system responds with, "There is no flight 94," rather than, "There is no aircraft for flight 94." We have taken care to produce appropriately rephrased verbal responses in such cases. With a yes-no question, the system makes two calls to the database, with the first determining the flight set for the topic, and the second determining the set after filtering based on the predicate is incorporated. An appropriate response can then be generated, considering both database tables.

The history data structure contains not only the frames associated with previously mentioned noun phrases and their modifiers, but also the previously displayed table, the previous state of the ticket under development, and previously booked tickets or first legs of a round trip ticket. The system frequently consults the tickets, as well as other elements from the history, to decide how to proceed with the dialogue. For example, users can refer to entries in the table, such as "the third one," and also to relative dates, such as "the day after tomorrow," requiring access to elements in the history data structure. A return date mentioned early in the dialogue is retained in the history for later reference. Return flights inherit all appropriate restrictions from the forward leg.

Dialogue The system can be operated in both a non-booking and a booking mode. In the latter, the system launches a reservations plan upon user request, which includes a number of subgoals initiated by either the system or the user. Once a user initiates a booking, a complex series of events transpires, in which the system is actively interpreting the state of the ticket and initiating both explicit requests to the user and calls to the database to provide relevant additional information. It also displays a facsimile of the ticket, and slots get filled in as they become specified. The system can carry the user all the way through a round trip booking, checking out that there are unique flight/fare/date specifications for both legs, and making sure that the dates are not violated by fare restrictions.

Figure 1 gives a block diagram of the control flow for managing discourse and dialogue. As shown in the figure, both the user and the system may issue questions to the back-end component. These questions are processed the same way, updating both the discourse and dialogue components accordingly. For instance, when the user has booked a particular flight but has said nothing about fares, the system can simply issue the request "Show fares" to the back-end. The discourse history will incorporate automatically the relevant flight information. If a user query is ambiguous, the system defers calling the database until it has queried the user for resolution of the ambiguity. After the system has answered the user's question, it assesses the dialogue state, which is maintained as a stack. When the dialogue stack is popped, the system may update the information contained in the ticket. The system may decide at this point to take the initiative, anticipating the user's needs.

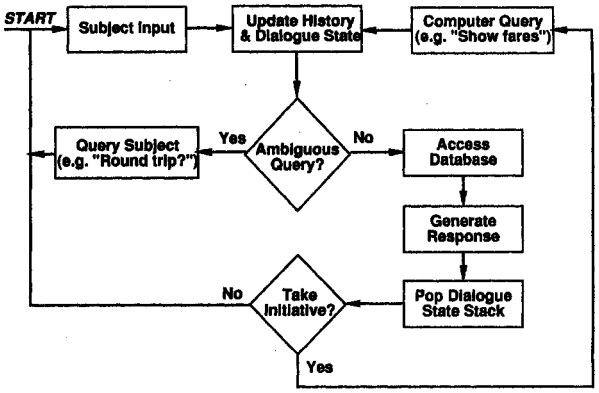

Figure 1: Block diagram of discourse/dialogue model.

DATA COLLECTION

As is the case with other efforts [3,2,1], our data are collected under simulation. Nevertheless, we wanted the simulation to reflect as much as possible the system that we are developing. In this section, we will briefly describe some design issues and document the actual collection process, contrasting it with the paradigm previously followed for ATIS data collection at TI. Further details can be found elsewhere [5].

Wizard *vs.* System By far the most important difference between the data collection procedures at TI and MIT is the way system simulation is conducted during data collection. TI made use of a "wizard" paradigm, in which a highly skilled experimenter interprets what was spoken, converts it into a form that enables database access, and produces an answer for the subject [3,2]. Based on our previous positive experience with collecting spontaneous speech for a different domain [10], we decided to explore an alternative paradigm that makes use of the system under development to do most of the work. Data collection is accomplished by having the experimenter, a fast and accurate typist, type verbatim to the system what was spoken, after removing spontaneous speech disfluencies. The actual interpretation and response generation is accomplished by the system without further human intervention. If the sentence cannot be understood by the system, an error message is produced to help the subject make appropriate modifications.

Displays Since the average traveller is not likely to be knowledgeable of the format and display of the Official Airline Guide (OAG), we have translated many of the cryptic symbols and abbreviations that OAG uses into easily recognizable words, to make the display tables more intelligible. In addition, we do not overload the subject with extraneous information not explicitly asked for, such as meals or aircraft for the flights.

System Feedback Our system provides explicit feedback to the subject in the form of text and synthetic speech, paraphrasing its understanding of the sentence. By providing confirmation to the subject of what was understood, the system greatly reduces the confusion and frustration that may arise due to a

misunderstanding between the system and the subject. In addition, a verbal response implicitly encourages the notion of human/machine interactive dialogue. Verbal responses are also given for system failures, which can occur in three ways: there could be unknown words, the sentence might not parse, or the back end processing might produce an error. Each of these carries a distinct error message communicating the nature of the problem to the subject.

COMPARATIVE ANALYSES

In this section we compare several variables measured on a subset of the data collected at MIT and TI, as shown in Table 1. Our higher yield is probably due to the fact that the system can respond much faster than a wizard; the process of translating the sentences into an NLParse command [2] by hand can sometimes be quite time-consuming. The speaking rate of the MIT sentences was more than 70% higher that that of the TI sentences. Filled pauses appear in 8.1% of the TI sentences, but only in 1.3% of the MIT sentences. Similarly, lexical false starts are twice as likely in the TI data as compared with the MIT data. Linguistic false starts are almost six times more likely to occur in the TI data as compared with the MIT data.

Variables	TI Data	MIT Data
# Sentences	774	1582
Yield (Utts. per Hour)	39	53
Ave. # Words/Second	1.18	2.04
% Utt. with Filled Pauses	8.1	1.3
% Utt. with Lexical False Starts	6.0	2.8
% Utt. with Linguistic False Starts	5.9	1.0

Table 1: Comparisons between data collected at TI and MIT.

Figure 2 compares the size of the lexicon as a function of the number of training sentences collected at TI and MIT. The vocabulary size grows at a much slower rate (about 20 words per 100 training sentences) for the MIT ATIS data than for the TI data (about 50 words per 100 training sentences). Also included on the figure for reference is a plot of the growth rate for our VOYAGER corpus, which was collected using the same paradigm as we have used for ATIS. A previous comparison of the TI data and the MIT VOYAGER data [4] led to the conclusion that the VOYAGER *domain* was intrinsically more restricted. Our newly collected data in the ATIS domain suggest that the data collection paradigm may be a more critical factor. Thus, one may argue that our data collection procedure is better able to encourage the subjects to stay within the domain.

EVALUATION AND DISCUSSION

We have now collected well over 4,500 utterances of data, and these have been extremely useful for helping us improve the design of the system. The system was able to handle about three fourths of the presented utterances on average. We are gradually expanding the grammar to cover forms that seem reasonable and within the domain. In addition, the discourse model has undergone significant improvements over time, as has the dialogue model.

In February, 1991, we evaluated the natural language and response generation components of our system using text input

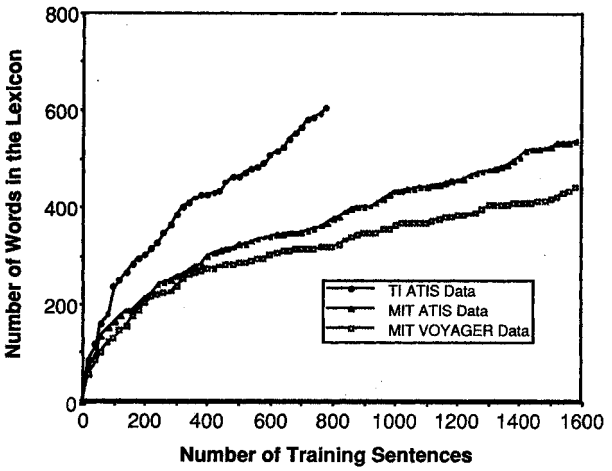

Figure 2: The size of the lexicon as a function of the number of training sentences for the TI and MIT ATIS training sets, as well as the MIT VOYAGER training set.

Data Set	No. of Sentences	Correct (%)	Incorrect (%)	No Answer (%)
TI	145	56.6	1.4	42.1
MIT	198	74.2	1.5	24.2

Table 2: Results for the TI and MIT ATIS test sets.

on a test set provided by the National Institute of Standards and Technology (NIST) and taken from the data collected at TI. We ran a comparable evaluation on a test set of utterances collected at MIT. Table 2 summarizes the results from this evaluation on what is known as Class A utterances, i.e., utterances that are context independent and within domain[2]. In both cases, the system was trained on data from both TI and MIT. For each test set, we provide the number of utterances in the set, and the percentages of utterances which produced correct answers, incorrect answers, and no answers. System performance was significantly higher on the MIT test set than on the TI test set.

We have recently begun a critical examination of the types of output our system provides to the user. System responses have in general been quite verbose, with all modifiers picked up from the history being intentionally mentioned. Now that the history model is more capable, we are beginning some experiments where the system's verbal responses are made to be more terse, explicitly mentioning only new information. Figure 3 shows an actual dialogue with a subject, where the original answers have been replaced by these less verbose ones. We feel that in most cases the table contains the necessary information missing from the verbal response. A possible subject for future experiments is whether users prefer a less verbose response.

A further examination of the data has indicated how subtle variations in the system's prompts can greatly affect the type of

[2]The "answers" for these utterances were obtained from the OAG database and provided in a standardized format called CAS (Common Answer Specification). These were compared against "reference" CAS's using a comparator provided by NIST.

```
USER: I'D LIKE A ROUND TRIP TICKET FROM DALLAS TO DENVER
ATIS: What date will you be travelling on?
USER: I'D LIKE TO LEAVE AUGUST FIFTEENTH
ATIS: Here are the round-trip flights from dallas to denver
      on Thursday August 15 <shows table>
USER: HOW MUCH IS THE CHEAPEST FLIGHT
ATIS: the cheapest flights. <Shows table>
USER: OKAY I'D LIKE TO GO ON UNITED
ATIS: Here are the united flights <Shows table>
USER: OKAY I'LL TAKE THE ONE AT ELEVEN TWENTY FIVE IN THE MORNING
ATIS: There are no flights leaving at 11:25 A.M. I'm sorry,
      you'll have to make another selection.
USER: I'D LIKE TO ARRIVE BEFORE NOON
ATIS: Here are the flights arriving before 12:00 P.M.
     Please select a booking. <Shows table>
USER: OKAY I'LL TAKE FLIGHT UNITED FOUR SEVENTY FIVE
ATIS: I'll update the ticket for you. I'll book united
      flight 475 for you. <System shows restriction information
      and warns user that restrictions require returning between
      Aug.18 and Sept.14. > Shall I show you the return flights?
USER  OKAY
ATIS: What date will you be travelling on?
USER: ON SEPTEMBER TWENTY SIXTH
ATIS: I'm sorry, but you must return before Saturday September 14.
      What date will you be travelling on?
USER: SEPTEMBER THIRTEENTH
ATIS: Here are the united flights from denver to dallas fort worth
      ap/80 class qw and qx on Friday September 13. Please select a
      booking. <User asks what "QW" means and system shows
      explanation table>
USER: I'LL TAKE THE LATEST FLIGHT
ATIS: I'll update the ticket for you. I'll book flight 664 for you.
      Can I help you with something else?
```

Figure 3: A dialogue between a subject and the system involving a round trip reservation

data observed. This is consistent with previous studies dealing with speech elicited over the telephone [9]. In our system, when the subject has successfully booked a forward leg, the system is prepared to display for them the set of applicable return flights. In our original implementation, the system asked the question, "Can I help you with the return flights?" at this point, expecting a cooperative yes/no answer. However, only 3 out of 18 subjects (17%) presented with this prompt responded with "Yes" or an equivalent (e.g., "Okay"). In contrast, a more explicit prompt, "Shall I show you the return flights?", produced a "Yes" or equivalent response in 23 of 27 subjects (85%).

CONCLUSION

Our focus on dialogue has enabled us to produce an interactive, mixed-initiative dialogue system. We have been successful in using this system in place of a wizard in data collection. We believe that this "system-assisted" paradigm offers several advantages. Since the system used to collect the data is under continual development, we can periodically replace it with an improved version, where the improvement will be guided by the data already collected. Furthermore, this approach to data collection provides relatively realistic sample data of human-machine interaction. This distinguishes it from wizard data, where the human wizard will answer any query that the back-end can answer. The data allow us to readily assess the subject's ability to adapt to the system and the system's ability to interact in an appropriate way. A potential disadvantage of this method of data collection is that the baseline system is constantly evolving. Data collected in an earlier session may differ significantly from data collected later. A possible consequence is that a dialogue may become incoherent at a later stage in system development.

Comparative analyses of data collected at TI and MIT reveal significant differences in many dimensions. Given the many ways in which the two procedures differ, it is not always easy to attribute the discrepancies to one single factor. The large difference in the frequency of spontaneous speech events leads us to believe that these effects may not be as prevalent in actual system usage as in a simulated mode.

ACKNOWLEDGEMENT

We would like to acknowledge the help of Claudia Sears and Christie Winterton, who served as experimenters and transcribers for much of the data, and Victoria Palay, who helped in the coordination of the data collection. We also wish to thank Bridget Bly of SRI for providing CAS reference answers for the MIT ATIS test set.

REFERENCES

[1] Bly, B., P. Price, S. Tepper, E. Jackson, and V. Abrash, "Designing the Human Machine Interface in the ATIS Domain," *Proc. Third Darpa Speech and Natural Language Workshop*, pp. 136-140, Hidden Valley, PA, June 1990.

[2] Hemphill, C. , J. Godfrey, and G. Doddington, "The ATIS Spoken Language System Pilot Corpus," *Proc. Third Darpa Speech and Natural Language Workshop*, pp. 96-101, Hidden Valley, PA, June 1990.

[3] Kowtko, J. and P. Price, "Data Collection and Analysis in the Air Travel Planning Domain," *Proc. Second Darpa Speech and Natural Language Workshop*, pp. 119-125, Harwichport, MA, October 1989.

[4] Norton, L., D. Dahl, D. McKay, L. Hirschman, M. Linebarger, D. Magerman, and C. Ball, "Management and Evaluation of Interactive Dialog in the Air Travel Domain," *Proc. Third Darpa Speech and Natural Language Workshop*, pp. 141-146, Hidden Valley, PA , June 1990.

[5] Polifroni, J., S. Seneff, and V. Zue, "Collection of Spontaneous Speech for the ATIS Domain and Comparative Analyses of Data Collected at MIT and TI," *Proc. Fourth Darpa Speech and Natural Language Workshop*, Pacific Grove, CA, February 1991.

[6] Price P., "Evaluation of Spoken Language Systems: The ATIS Domain," *Proc. Third Darpa Speech and Natural Language Workshop*, pp. 91-95, Hidden Valley, PA , June 1990.

[7] Seneff, S., J. Glass, D. Goddeau, D. Goodine, L. Hirschman, H Leung, M. Phillips, J. Polifroni and V. Zue, "Development and Preliminary Evaluation of the MIT ATIS System," *Proc. Fourth Darpa Speech and Natural Language Workshop*, Pacific Grove CA, February 1991.

[8] Seneff, S., L. Hirschman, and V. Zue, "Interactive Problem Solving and Dialogue in the ATIS Domain," *Proc. Fourth Darpa Speech and Natural Language Workshop*, Pacific Grove, CA February 1991.

[9] Spitz, J. "Collection and Analysis of Data from Real Users: Implications for Speech Recognition/Understanding Systems" *Proc Fourth Darpa Speech and Natural Language Workshop*, Pacific Grove, CA, February 1991.

[10] Zue, V., N. Daly, J. Glass, H. Leung, M. Phillips, J. Polifroni, S Seneff, and M. Soclof, "The Collection and Preliminary Analysis of a Spontaneous Speech Database," *Proc. Second Darpa Speech and Natural Language Workshop*, pp. 126-134, Harwichport, MA October 1989.