Evaluation Results for the Talk'n'Travel System

David Stallard
BBN Technologies, Verizon
70 Fawcett. St.
Cambridge, MA, 02140
Stallard@bbn.com

ABSTRACT

We describe and present evaluation results for Talk'n'Travel, a spoken dialogue language system for making air travel plans over the telephone. Talk'n'Travel is a fully conversational, mixed-initiative system that allows the user to specify the constraints on his travel plan in arbitrary order, ask questions, etc., in general spoken English. The system was independently evaluated as part of the DARPA Communicator program and achieved a high success rate.

.

1. INTRODUCTION

This paper describes and presents evaluation results for Talk'n'Travel, a spoken language dialogue system for making complex air travel plans over the telephone. Talk'n'Travel is a research prototype system sponsored under the DARPA Communicator program (MITRE, 1999). Some other systems in the program are Ward and Pellom (1999), Seneff and Polifroni (2000) and Rudnicky et al (1999). The common task of this program is a mixed-initiative dialogue over the telephone, in which the user plans a multi-city trip by air, including all flights, hotels, and rental cars, all in conversational English over the telephone. A similar research program is the European ARISE project (Den Os et al, 1999).

An earlier version of Talk'n'Travel was presented in (Stallard, 2000). The present paper presents and discusses results of an independent evaluation of Talk'n'Travel, recently conducted as part of the DARPA Communicator program.

The next section gives a brief overview of the system.

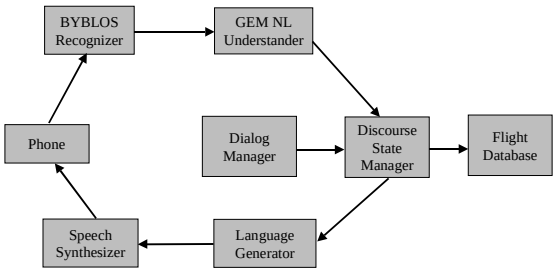
2. SYSTEM OVERVIEW

The figure shows a block diagram of Talk'n'Travel. Spoken language understanding is provided by statistical N-gram speech recognition and a robust language understanding component. A plan-based dialogue manager coordinates interaction with the user, handling unexpected user input more flexibly than conventional finite-state dialogue control networks. It works in tandem with a state management component that adjusts the current model of user intention based on the user's last utterance in context.

Meaning and task state are represented by the path constraint representation (Stallard, 2000). An inference component is included which allows the system to deduce implicit requirements from explicit statements by the user, and to retract them if the premises change.

The system is interfaced to the Yahoo/Travelocity flight schedule website, for access to live flight schedule information. Queries to the website are spawned off in a separate thread, which the dialogue manager monitors ands reports on to the user.

Figure 1 : System Architecture



3. DIALOGUE STRATEGY

Talk'n'Travel employs both open-ended and directed prompts. Sessions begin with open prompts like "What trip would you to take?". The system then goes to directed prompts to get any information the user did not provide ("What day are you leaving?", etc). The user may give arbitrary information at any prompt, however. The system provides implicit confirmation of the change in task state caused by the user's last utterance ("Flying from Boston to Denver tomorrow ") to ensure mutual understanding.

The system seeks explicit confirmation in certain cases, for example where the user appears to be making a change in date of travel. Once sufficient information is obtained, the system offers a set of candidate flights, one at a time, for the user to accept or reject.

4. EVALUATION

4.1 Evaluation Design

The 9 groups funded by the Communicator program (ATT, BBN, CMU, Lucent, MIT, MITRE, SRI, and University of Colorado)

took part in an experimental common evaluation conducted by the National Institute of Standards and Technology (NIST) in June and July of 2000. A pool of approximately 80 subjects was recruited from around the United States. The only requirements were that the subjects be native speakers of American English and have Internet access. Only wireline or home cordless phones were allowed.

The subjects were given a set of travel planning scenarios to attempt. There were 7 such prescribed scenarios and 2 open ones, in which the subject was allowed to propose his own task. Prescribed scenarios were given in a tabular format. An example scenario would be a round-trip flight between two cities, departing and returning on given dates, with specific arrival or departure time preferences.

Each subject called each system once and attempted to work through a single scenario; the design of the experiment attempted to balance the distributions of scenarios and users across the systems.

Following each scenario attempt, subjects filled out a Web-based questionnaire to determine whether subjects thought they had completed their task, how satisfied they were with using the system, and so forth. The overall form of this evaluation was thus similar to that conducted under the ARISE program (Den Os, et al 1999).

4.2 Results

Table 1 shows the result of these user surveys for Talk'n'Travel. The columns represent specific questions on the user survey. The first column represents the user's judgement as to whether or not he completed his task. The remaining columns, labeled Q1-Q5, are Likert scale items, for which a value of 1 signifies complete agreement, and 5 signifies complete disagreement. Lower numbers for these columns are thus better scores. The legend below the table identifies the questions.

Table 1 : Survey Results

	Task Comp%	Q1	Q2	Q3	Q4	Q5
BBN	80.5%	2.23	2.09	2.10	2.36	2.84
Mean	62.0%	2.88	2.23	2.54	2.95	3.36

Scale: 1 = strongly agree, 5 = strongly disagree

Q1 It was easy to get the information I wanted
Q2 I found it easy to understand what the system said
Q3 I knew what I could do or say at each point in the dialog
Q4 The system worked the way I expected it to
Q5 I would use this system regularly to get travel information

The first row gives the mean value for the measurements over all 78 sessions with Talk'n'Travel. The second row gives the mean value of the same measurements for all 9 systems participating.

Talk'n'Travel's task completion score of 80.5% was the highest for all 9 participating systems. Its score on question Q5, representing user satisfaction, was the second highest.

An independent analysis of task completion was also performed by comparing the logs of the session with the scenario given. Table 2 shows Talk'n'Travel's results for this metric, which are close to that seen for the user questionnaire.

Table 2: Objective Analysis

Completion of required scenario	70.5%
Completion of different scenario	11.5%
Total completed scenarios	82.0%

Besides task completion, other measurements were made of system operation. These included time to completion, word error rate, and interpretation accuracy. The values of these measurements are given in Table 3.

Table 3: Other Metrics

Average time to completion	246 secs
Average word error rate	21%
Semantic error rate/utterance	10%

4.3 Analysis and Discussion

We analyzed the log files of the 29.5% of the sessions that did not result in the completion of the required scenario. Table 4 gives a breakdown of the causes.

Table 4: Causes of Failure

City not in lexicon	39% (9)
Unrepaired recognition error	22% (5)
User error	17% (4)
System diversion	13% (3)
Other	9% (2)

The largest cause (39%) was the inability of the system to recognize a city referred to by the user, simply because that city was absent from the recognizer language model or language understander's lexicon. These cases were generally trivial to fix.

The second, and most serious, cause (22%) was recognition errors that the user either did not attempt to repair or did not succeed in repairing. Dates proved troublesome in this regard, in which one date would be misrecognized for another, e.g. "October twenty third" for "October twenty first

Another class of errors were caused by the user, in that he either gave the system different information than was prescribed by the scenario, or failed to supply the information he was supposed to. A handful of sessions failed because of additional causes, including system crashes and backend failure.

Both time to completion and semantic error rate were affected by scenarios that failed because because of a missing city. In such scenarios, users would frequently repeat themselves many times in a vain attempt to be understood, thus increasing total utterance count and utterance error.

An interesting result is that task success did not depend too strongly on word error rate. Even successful scenarios had an average WER of 18%, while failed scenarios had average WER of only 22%.

A key issue in this experiment was whether users would actually interact with the system conversationally, or would respond only to directive prompts. For the first three sessions, we experimented with a highly general open prompt ("How can I help you?'), but quickly found that it tended to elicit overly general and uninformative responses (e.g. "I want to plan a trip"). We therefore switched to the more purposeful "What trip would you like to take?" for the remainder of the evaluation. Fully 70% of the time, users replied informatively to this prompt, supplying utterances "I would like an American flight from Miami to Sydney" that moved the dialogue forward.

In spite of the generally high rate of success with open prompts, there was a pronounced reluctance by some users to take the initiative, leading them to not state all the constraints they had in mind. Examples included requirements on airline or arrival time. In fully 20% of all sessions, users refused multiple flights in a row, holding out for one that met a particular unstated requirement. The user could have stated this requirement explicitly, but chose not to, perhaps underestimating what the system could do. This had the effect of lengthening total interaction time with the system.

4.4 Possible Improvements

Several possible reasons for this behavior on the part of users come to mind, and point the way to future improvements. The synthesized speech was fairly robotic in quality, which naturally tended to make the system sound less capable. The prompts themselves were not sufficiently variable, and were often repeated verbatim when a reprompt was necessary. Finally, the system's dialogue strategy needs be modified to detect when more initiative is needed from the user, and cajole him with open prompts accordingly.

5. ACKNOWLEDGMENTS

This work was sponsored by DARPA and monitored by SPAWAR Systems Center under Contract No. N66001-99-D-8615.

6. REFERENCES

[1] MITRE (1999) DARPA Communicator homepage http://fofoca.mitre.org/

[2] Ward W., and Pellom, B. (1999) The CU Communicator System. In *1999 IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado.

[3] Den Os, E, Boves, L., Lamel, L, and Baggia, P. (1999) Overview of the ARISE Project. *Proceedings of Eurospeech, 1999*, Vol 4, pp. 1527-1530.

[4] Miller S. (1998) The Generative Extraction Model. Unpublished manuscript.

[5] Constantinides P., Hansma S., Tchou C. and Rudnicky, A. (1999) A schema-based approach to dialog control. *Proceedings of ICSLP*, Paper 637.

[6] Rudnicky A., Thayer, E., Constantinides P., Tchou C., Shern, R., Lenzo K., Xu W., Oh A. (1999) Creating natural dialogs in the Carnegie Mellon Communicator system. *Proceedings of Eurospeech, 1999*, Vol 4, pp. 1531-1534

[7] Rudnicky A., and Xu W. (1999) An agenda-based dialog management architecture for soken language systems. In *1999 IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado.

[8] Seneff S., and Polifroni, J. (2000) Dialogue Management in the Mercury Flight Reservation System. *ANLP Conversational Systems Workshop.*