ICSLP 2002 (In Press)

Analysis of User Behavior under Error Conditions in Spoken Dialogs

Jongho Shin, Shrikanth Narayanan, Laurie Gerber, Abe Kazemzadeh, Dani Byrd*

University of Southern California – Integrated Media Systems Center
Speech Analysis and Interpretation Laboratory: http://sail.usc.edu

ABSTRACT

We focus on developing an account of user behavior under error conditions, working with annotated data from real human-machine mixed initiative dialogs. In particular, we examine categories of error perception, user behavior under error, effect of user strategies on error recovery, and the role of user initiative in error situations. A conditional probability model smoothed by weighted ASR error rate is proposed. Results show that users discovering errors through implicit confirmations are less likely to get back on track (or succeed) and take a longer time in doing so than other forms of error discovery such as system reject and reprompts. Further successful user error-recovery strategies included more rephrasing, less contradicting, and a tendency to terminate error episodes (cancel and startover) than to attempt at repairing a chain of errors.

1. INTRODUCTION

Modeling human-machine spoken dialog interactions is gaining a lot of attention [4,5] with the recent deployment of several complex dialog systems, for e.g., [1,2,3]. An important aspect of this problem is the understanding and modeling user behavior to enable realistic optimization of dialog strategies. It is well known that many of the underlying components of the state-of-the-art dialog systems such as automatic speech recognition and understanding rely on data-driven statistical models and, in general, are prone to errors of varying types and extent. In addition, there are other possible systems and user induced errors. Our work targets user behavior modeling under such error conditions in the context of human-machine spoken dialogs.

The DARPA Communicator spoken dialog systems, implemented at several sites, represent some of the most recent advances in the design of mixed-initiative spoken language systems [1,2,5]. The availability of transcripts of realistic spoken dialogs from some of those systems provides an excellent opportunity to investigate the behavior of human and machine interactions in mixed-initiative dialogs. In the present work we set out to understand the dynamics of user behavior under system errors and how the combination of system errors and user reactions to them affect the ultimate success of a dialog. In preparation for this study, we annotated a portion of the June 2000 Communicator dialogs for several features, including a categorization of both user and system behavior. The data and the extended annotation scheme are described in section 2. The results of our study are described in section 3. The paper concludes with a summary and discussion of the results in section 4.

2. DATA AND ANNOTATION

The data used were the orthographically-transcribed travel arrangement dialogs from the DARPA Communicator project recorded in June 2000. Each dialog consists of some number of exchanges between a computer travel agent and a human and is represented as a three-line triple consisting of a system utterance, a user utterance (manually transcribed from recordings), and what the ASR system heard and provided as input to the dialog system.

The data and the collection procedure are described in detail in [5]. In the Communicator dialogs, 85 experimental subjects interacted with 9 different "travel agent" systems. Of the 765 possible dialogs, many are empty, or contain no user participation. We worked with about 141 of those total dialogs (that consisted of at least 1 turn). The average length of these dialogues was 18 exchanges. The amount of data is comparable to the data considered in a similar study by Aberdeen et al [6].

2.1. Tagging

Following a review of the recent work on analysis of human computer dialogs, we devised a tagging scheme consisting of 23 tags with which to monitor 3 dimensions of the dialogs: user behavior, system behavior, and task status. Since our goal was to do a quantitative analysis of the (disruptive) effect of errors, existing tagging schemes, while instructive, were not directly applicable. Automatic analysis of error conditions beyond the ASR word error rate is difficult without the aid of manual tagging. Hence, manual tagging was necessary. However, for example, unlike [6], we do not keep track of the subtask in which the error occurred, nor do we distinguish between dialog acts as in [7]. Finally, the user utterances in the communicator data are very short, averaging 3 words. Under these circumstances, we also have not made an to attempt labeling disfluencies as projects dealing with longer, more open-ended utterances have done [8] [9][10].

The detailed tag set together with usage conventions and examples of application are provided in http://sail.usc.edu/dialog/model_tags. Briefly, the tag set for our purposes included (1) SYSTEM tags: explicit confirmation, implicit confirmation, help, system repeat, reject, non sequitur (2) USER tags: repeat, rephrase, contradict, frustrated, change request, startover, scratch, clarify, acquiesce, hang-up (3) TASK tags: error (at the recognized utterance), back on track, task success.

For error segments, we locate the beginnings of errors, and place a generic "error" tag on the ASR output that resulted in an error (Note that the standard ASR word error rate for each turn is also calculated). Within error segments we focus on three phenomena: system utterances which exhibit a system reaction to the error, user utterances which react to or try to correct the error, and the means by which the user becomes aware of the error. Sometimes the user becomes aware of an error because of a system rejection such as, *"I'm sorry, I couldn't understand you."* or a verbatim repetition of a system prompt for information. Other times implicit confirmations or non sequiturs in system utterances alert the user to the presence of an error, in which case the user must try to make the system aware of the error. Because the scenarios were conducted by paid subjects arranging for hypothetical travel for this particular data collection, some users had a tendency to acquiesce to errors that proved

difficult to correct, or even to change the nature of the travel request in response to repeated recognition errors. These deviations from the original plan are also marked.

Finally, we tag the point at which the dialog gets back-on-track (BOT), marking the system utterance in which the user could reasonably discover that the portion of the task derailed by an error has been successfully understood. At the end of the dialog we indicate whether the arrangements were successfully completed or ended in a hang-up or acquiescence to some error. The tagging was done by two annotators and showed 87% inter-annotator agreement. The tagging conventions used allow the assignment of all applicable tags to the dialogs. The agreement measure used was the number of identically tagged lines, divided by the number of lines reviewed and tagged. The measure is conservative in that it counts as agreement cases where 100% identical tagging appears on exactly the same line for both annotators. It does not include partial overlap, or positional offset.

Following the tagging itself, we analyzed the dialogs and user histories from several perspectives, seeking patterns in user behavior, and correlations between user behavior and the length and severity of error segments.

3. RESULTS AND DISCUSSION

Firstly it is useful to get a general sense of the presence of errors in the dialogs. The data, overall, is dominated by errors of various types. The roughly 2528 turns we tagged consists of 141 dialogs conducted with 35 paid subjects. The dialogs contain 235 error segments. Note that according to our definition an error segment can (1) end in either by getting back on track (BOT) with perhaps a complete success, acquiescence or abort (2) be nested within another error segment. Of these 235 segments, 78% got back on track.

Figure 1 provides the distribution of error segment length (number of turns) in the data. About 80% of these are between 1-9 turns with most of them between 2 to 4 turns. Of these, the average length of the error segments that eventually get back-on-track is 6.7 and those that never recover is 10. From these numbers alone, we do not know whether the length of the unrecovered errors represents something about the system or user, or if it represents some threshold of user tolerance for error resolution beyond which users will simply hang up rather than continue.

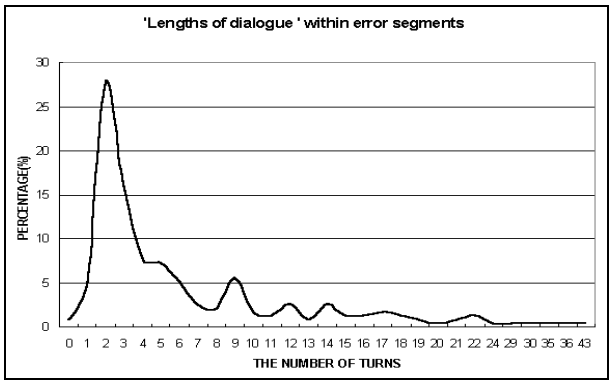

Figure 1: Normalized histogram of the length of error segments (number of turns).

We present analysis results on the following points: (1) Categories of error perception (2) User behavior under error including user initiative in error vs. non-error situations

3.1. Categories of error perception

Here, we see whether the manner in which the user discovered the error affects the time to get back on track. In the case of a system prompt repetition or a system rejection, the user is explicitly made aware of an "error" (from its perspective). In the case of an implicit confirmation or a system non sequitur, it is up to the user to notice that an error has occurred and draw the system's attention to this. In Table 1, we present error segments grouped by the way in which the user becomes aware of the error, to see if the way in which the error is discovered affects the time to recover or success in recovery.

We can roughly divide the error discovery types into high frequency (system rejection, implicit confirmation, & system prompt repeat), and low frequency (explicit confirmation & non-sequitur). Among the high-frequency error discovery types, it is striking that *implicit* confirmation results in a much longer time to get back on track (10 exchanges vs. 6), and a much lower rate of getting back-on-track at 68%, compared to 80% and 90% for the other high-frequency errors.

Error perception	# of err segments	avg err length for BOT	avg err length not BOT	%BOT
Reject	35	6	7.8	83%
Implicit	25	9.6	14.6	68%
Repeat	21	5.8	13	90%
Explicit	10	5.5	8.75	60%
Non-seq	9	6	7.5	77%

Table 1: Lengths of error segments which did get back-on-track (BOT) and those which didn't, as well as the percentage of errors that eventually got back on track.

3.2. User behavior under error

We next examine the distribution of user behaviors in coping with errors. Figure 2 shows the distribution on the user behavior immediately following an error (in the previous turn).

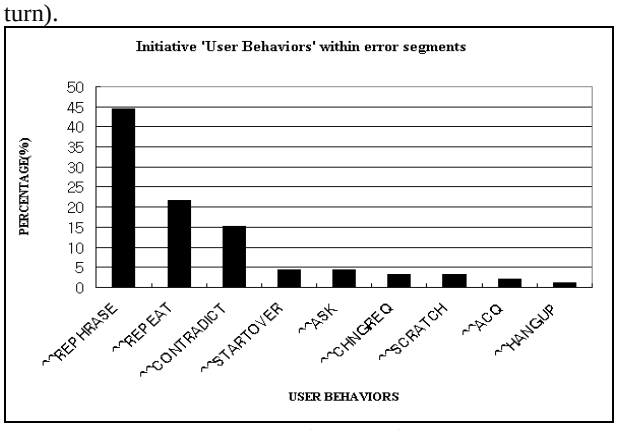

Figure 2: 'User Behavior' after the first error within an error segment. Rephrasing was the most frequent user behavior and Hang-up was the least frequent user behavior.

The next two tables show the distribution of user strategies for segments that eventually did get back on track and for those that never got back on track:

frequency normalized for length of errors	User strategy in Errors that got back-on-track
0.130	Repeat
0.117	Rephrase
0.077	Contradict system
0.055	Start over
0.045	Ask
0.022	*Change request*
0.015	Scratch
0.005	Acquiesce to error

Table 2: Prevalence of user strategies in error segments which eventually got back on track.

We observe that users in the successful error recoveries (see Table 2) use significantly ($p < 0.1$, ANOVA) more rephrasing than those in the unrecovered errors and less contradictions (e.g. "not 3 am, 3 pm") (Table 3). They also make use of the "start over" and "scratch" features more to terminate error episodes rather than trying to repair chains of errors. Users in successful error recoveries were also much more likely to work around system weaknesses by changing their travel plans. While this apparently got the dialog back on track, it is not a viable strategy for real travel arrangements.

frequency normalized for length of errors	User strategy in Non-back-on-track
0.114	Repeat
0.102	Contradict system
0.071	Rephrase
0.055	*Hang up*
0.031	Start over
0.024	Ask
0.012	Scratch
0.012	Acquiesce to error
0.004	Change request

Table 3: Prevalence of user strategies in error segments which did not get back-on-track

Degree of Error and User behavior: Errors in spoken dialogs are not merely binary valued and it is critical to incorporate the degree of error into the modeling. To illuminate user behavior under error further, we considered the user response conditioned on the system strategy to estimate the probability P(User Behavior| System Behavior), P(U|S) from now on. It has been well accepted in the field that ASR word error rate (WER) is a good correlate of dialog performance [5]. Hence as a first approximation, we smoothed the probability mass of P(U|S) using an exponentially-weighted WER measure (1-10**(-WER*k/100)) that maps WER (which can be between 0 and infinity) to a range between 0 and 1. For the calculations below we chose k=1; it could vary from system to system. The results are shown in Figure 3. The most common user behavior here is rephrasing or repeating the previous request, contributing to 82% of all user responses under error. Canceling/changing the previous request or starting over are relatively rare user behaviors under error.

This is further exemplified in Figure 4 that shows the conditional (smoothed) distribution for P(U|S=SYSTEM REPEAT), corresponding to a highly popular system strategy when the system is "cognizant" of an error.

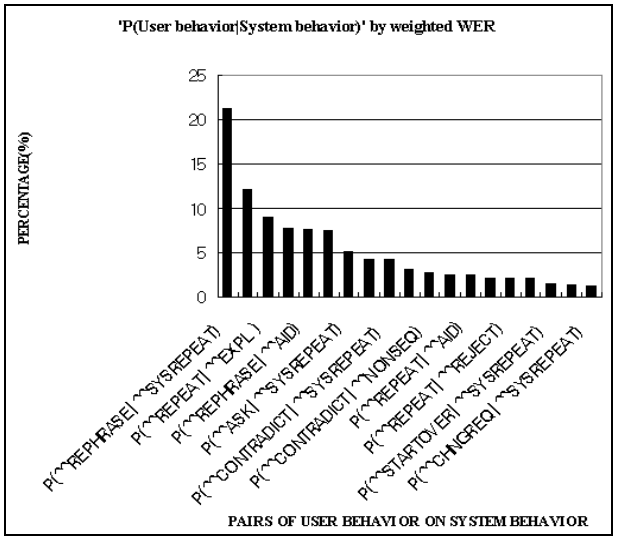

Figure 3: P (User behavior | System behavior) smoothed by exponentially weighted WER.

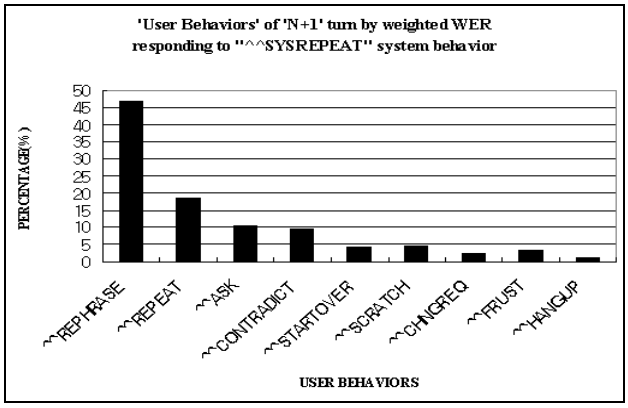

Figure 4: Smoothed Conditional Probability of User Behavior in (N+1)th turn based on weighted WER of 'SYSTEM REPEAT' system behavior in the N-th turn.

It is similarly interesting to look at user behavior when the system is not (necessarily) cognizant of an error such as when using an implicit confirmation strategy. Figure 5 shows the smoothed distribution for P(U|S=IMPLICIT). Not surprisingly, the user is most likely to contradict the erroneous system behavior.

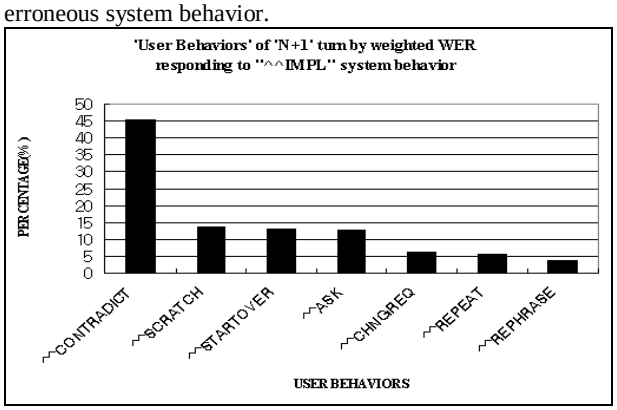

Figure 5: Smoothed Conditional Probability for User Behavior in (N+1)th turn based on weighted WER of 'IMPLICIT CONFIRM' system behavior in the N-th turn..

User initiative in error and non-error environments: Here we look at the user's tendency to use initiative over the course of the dialog. We have considered user initiative to be the cases where the user did not simply respond to system prompts, but attempted to guide the dialog themselves. The one part of the dialog that often looks the most like user initiative (and which often fails) is the response to the open prompt at the beginning of most of the dialogs. However, since this is a free-form answer to an open question, we have not tagged it as initiative. It is clear from Table 4 that user initiative behavior is significantly more in error segments than not ($p < 0.05$).

User Initiative tag	Frequency in error segments	Frequency in non-error segments
Ask	0.0319	0.0060
Contradict	0.0707	0.0121
General initiative	0.1647	0.0424

Table4: Frequency is normalized over all dialogs

4. SUMMARY

Modeling user behavior is one of the most challenging problems in spoken dialog systems research. Empirical analysis and modeling using real user data helps to illuminate user behavior patterns. The analysis reported represents a preliminary attempt at understanding user behavior under error and uncertainty in spoken dialogs.

Results show that users discovering errors through implicit confirmations are less likely to get back on track (or succeed) and take a longer time in doing so than other forms of error discovery such as system reject and reprompts. Further successful user error-recovery strategies included more rephrasing, less contradicting, and a tendency to terminate error episodes (cancel and startover) than to attempt at repairing a chain of errors.

The most frequent user behavior to get back on track from error segments when the system signals errors is to "rephrase" and "repeat". When a user discovers an error, say through an implicit confirmation, the user tends to "contradict" or "cancel" the action rather than "rephrase" and "repeat".

There are many open and confounding issues. One key issue relates to incorporating user behavior priors (i.e., probabilities) in the model. For example, we observe that some users seem better able to avoid and/or get out of trouble. The authors of [1] observe that in this specific experimental setup, where the subjects were paid participants with no real stake in successful task completion, some users were simply inattentive or careless. In the process of tagging the transcribed data, we additionally observed that some participants had much more trouble than others getting usable ASR output. Table 5 looks at some users who participated in 5 or more scenarios.

In Table 5, two users, A and B, seem particularly successful. Although they appear to have higher numbers of errors per dialog, this is probably because they did not give up, since they also have the highest rates of recovery with relatively short error episodes. Two other users, C and D, seem the least successful. D has a very low percentage of back-on-track errors, and C seems to experience inordinately long error episodes. When we looked at the strategies these users adopted under error we found that all users tried repeating themselves. However, the less successful users frequently hung up on the dialog or started the dialog sequence over; something that the successful users were less likely to do.

User ID	# of dials	Errs / dial.	%BOT	Avg length of err segmt
1	9	1.4	.69	8.5
2	9	1.4	.76	8.9
A	8	2.9	.87	7.8
B	8	2.4	.74	4.9
C	5	1.0	.60	10.2
D	5	1.4	.42	6.0

Table5: Error-proneness in users: % BOT is the percentage of error episodes that got back on track.

These types of prior user information need to be learnt and incorporated into the models. Ongoing work focuses on those questions and how a user model interacts with a system model in an optimization framework.

5. REFERENCES

[1] Ward, W., Pellom, B, "The CU communicator System", IEEE ASRU, pp. 341-344, 1999.

[2] Levin, E., Narayanan, S., Pieraccini, R., Biatov, K., Bocchieri, E., Di Fabbrizio, G., Eckert, W., Lee, S., Pokrovsky, A., Rahim, M., Ruscitti, P., and Walker, M *The AT&T-DARPA Communicator mixed-initiative spoken dialog system*, Proc. of ICSLP, (Beijing, China), pp. 122-125, 2000.

[3] Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., and Hetherington, L., "JUPITER: A Telephone-based Conversational Interface for Weather Information", IEEE Trans. Speech and Audio Proc., pp. 85-96, 2000.

[4] E. Levin, R. Pieraccini, W. Eckert, "A Stochastic Model of human-machine interaction for learning dialog strategies", IEEE Trans. Speech and Audio Proc., pp. 11-23, 2000.

[5] Walker, M., Aberdeen, J., Boland, J., Bratt, E., Garofolo, J., Hirschman, L., Le, A., Lee, S., Narayanan, S., Papineni, K., Pellom, B., Polifroni, J., Potamianos, A., Prabhu, P., Rudnicky, A., Sanders, G., Seneff, S., Stallard, D., and Whittaker, S., "DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection"., Proc. Eurospeech 2001.

[6] Aberdeen, J., Doran, C., Damianos, L., Bayer, S., and Hirschman, L., "Finding Errors Automatically in Semantically Tagged Dialogues", in *Proceedings of HLT*\. 173-178. 2001.

[7] Walker, M. and Passoneau, R. "Dialog Act Tags as Qualitative Dialog Metrics for Spoken Dialog Systems". in *Proceedings of, HLT 2001*, 2001.

[8] Levow, G.-A. "Characterizing and recognizing spoken corrections in human-computer dialogue." In *Proceedings of COLING/ACL-98*, 1998.

[9] Allen, J., and Core, M. "Draft of DAMSL: Dialog Act Markup in Several Layers". October, 1997.

[10] Langkilde, I, Walker, M., Wright, J., Gorin, A., Litman, D., "Automatic Prediction of Problematic Human-Computer Dialogues in "How May I Help You". In ASRU, 1999.

* Shri Narayanan was an investigator in the AT&T DARPA Communicator project during the June 2000 collection. The authors are grateful to the DARPA Communicator team participants for sharing the data.