Characterizing Task-Oriented Dialog using a Simulated ASR Channel

Jason D. Williams and Steve Young

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, United Kingdom
{jdw30,sjy}@eng.cam.ac.uk

Abstract

We describe a data collection consisting of task-oriented human-human conversations in a simulated ASR channel in which the WER is systematically varied. We find that users infrequently give a direct indication of having been misunderstood; levels of expert "initiative" increase with WER primarily due to increased grounding activity; and asking task-related questions appears to be a more successful repair strategy at moderate WER levels. A PARADISE analysis finds task completion most predictive of user satisfaction; efficiency is also important at lower WERs.

1. Introduction

Our chief research goal is applying machine-learning approaches to the dialog management component of spoken dialog systems (SDSs). This pursuit focuses on creating statistical models of users, based on data. In [1], we describe a dialog collection framework which is suitable for this purpose. We are also interested in gaining basic insights into conversational behavior in the automated speech recognition (ASR) channel at different word-error rate (WER) target levels, and our collection framework is also well-suited to this purposes. In particular, we hope to: show successful repair patterns relevant for SDSs, such as entering into dialog repair vs. asking task-related questions; characterize "natural" behavior, such as "initiative"; and identify appropriate reward measures which are likely to predict user satisfaction.

This paper has two goals. First, we aim to describe the experimental procedure used to collect the SACTI-1 corpus. SACTI stands for *Simulated ASR-Channel: Tourist Information*. Second, we seek to show patterns of user and wizard behavior in this corpus relevant to SDSs. The paper is organized as follows. Section 2 summarizes the data collection framework and discusses related work. Section 3 covers experimental procedure and annotation. Section 4 provides experimental results and discussion.

2. Background and Motivation

The data collection framework is described in detail in [1] and briefly summarized here. Two subjects, the "wizard" and "user," interact using an audio-only interface.

Both the wizard and user speech is segmented into utterances using an energy-based end-pointer. A turn-taking model similar to that commonly found in speech recognition systems is used, in which the user can interrupt or "barge-in over" the wizard, but not vice-versa. When the system is not listening to a speaker, that speaker hears a tic-toc sound.

When the wizard speaks, the user can hear the wizard directly. However, both the user and wizard are told the user is speaking to a speech recognizer, which places its interpretation of the user's speech on a screen in front of the wizard. In reality, the user is speaking to a typist who transcribes the user's speech. This transcription is then passed to a system which simulates speech recognition errors in text. By varying parameters in this confusion system, we can reliably simulate WERs from ~0% to very high levels.

The ASR simulation uses a phonetic confusion model and a language model to simulate recognition errors. The phonetic confusion model was trained on the TIMIT corpus using a monophone HMM. The language model was a bigram model, trained on a small corpus collected from earlier trial dialogs using the system and augmented with hand-crafted classes.

Work in [2] studies a variety of dialog phenomena in goal-directed Human-Computer (HC) and Human-Human (HH) conversation in the air-travel/transportation domain. They found that computer experts were more verbose than human experts, using on average 17-33 words/turn compared to humans' 10.1 words/turn. Conversely, users were much less verbose in HC conversation, using 2.8-4.8 words/turn in HC conversation and 7.2 words/turn in HH conversation. They also found that dialog initiative – i.e., dialog control – while difficult to reliably identify, was approximately evenly distributed in HH conversation but remained with the computer in approximately 90% of turns in HC conversation.

Other research [3] undertakes a study similar to the one presented here, differing in that it used a real speech recognizer. Also, the experiments in [3] used tasks in which the wizard primarily gave instructions to users. The tasks in this work were varied in nature.

3. Experimental procedure

We wanted to explore a variety of dialog genres, including information-seeking tasks, "Map tasks," and negotiations. After exploring a few domains, we found that the tourist information domain met our requirements.

The wizard was given a host of information about a fictitious town, and the user was given a task to complete. Example tasks included finding a hotel that meets a number of criteria, or planning a day of activities. The user was given a simple (accurate) map of the town, and the wizard was given a more detailed map. Some tasks had ambiguous solutions. A total of 24 tasks were created.

Each wizard interacted with 3 users, and each user undertook 4 tasks. Each wizard and user was greeted separately and their task explained using the same script. Subjects were a mixture of native and non-native speakers. Dialogs were allowed to run until the user ended the dialog, up to 10 minutes (but in some cases longer). At the end of each task, both the user and wizard were asked a similar set of 6 Likert-scored questions about their experiences covering task completion, speech recognition accuracy, ease of use, helpfulness of wizard, perceived ease of use for the other subject, and overall satisfaction.

Later, the orthographic transcription was completed by transcribing the wizard's utterances using a subset of the LDC conventions [4] with the LDC AG tool [5].

The simulation's state machine enforces a turn-taking regime in which just one participant can speak at one time. However, end-pointing errors led to states in which the end-pointer gave a subject the channel but they did not speak.

Thus, we segmented turns into the longest sequence of states in which only one participant was speaking.

Similar to [3], we were interested to assess how the wizard interpreted the previous user turn. Each wizard turn was classified into one "understanding category" (Table 1).

In addition, each turn was labeled with one or more temporally-ordered set of tags showing grounding behavior. These tags were inspired by the "Grounding Acts" of the finite state grounding model of [7], but modified in several respects. First, we annotated only acts which could be associated with text in the transcription -- for example, we annotated *ExplAck* only when it appeared as "Ok" or "I see" or similar, but not when implicit. Second, we added several acts for grounding behaviors which appeared frequently in the corpus, such as stating an interpretation of the other speaker's meaning/intention. Finally, we added two "content" acts – *Request* and *Inform*. *See Table 2*. Annotation was performed with ANVIL [6].

Label	Wizard's understanding of previous user turn
Full	All intentions understood correctly.
Partial	Some intentions understood; none misunderstood.
Non	Wizard made no guess at user intention.
Flagged-Mis	The wizard formed an incorrect hypothesis of the user's meaning, and signaled a dialog problem
Un-Flagged-Mis	The wizard formed an incorrect hypothesis of the user's meaning, accepted it as correct and continued with the dialog.

Table 1: Wizard understanding status categories

Tag	Meaning
Request	Question/request requiring response
Inform	Statement/provision of task information
Greet-Farewell	"Hello", "How can I help," "that's all", "Thanks", "Goodbye", etc.
ExplAck	Explicit statement of acknowledgement, showing speaker understanding of OS
Unsolicited-Affirm	Explicit statement of acknowledgement, showing OS understands speaker
HoldFloor	Explicit request for OS to wait
ReqRepeat	Request for OS to repeat their last turn
ReqAck	Request for OS to show understanding
RspAffirm	Affirmative response to *ReqAck*
RspNegate	Negative response to *ReqAck*
StateInterp	A statement of intention of OS
DisAck	Show of lack of understanding of OS
RejOther	Display of lack of understanding of speaker's intention or desire by OS

Table 2: Tag set. OS refers to "Other speaker."

4. Results and Discussion

We collected a total of 144 dialogs at four different WER levels (3). As WER increases, grounding behaviors become increasingly prevalent – Figure 1 shows the distribution of tags in wizard turns across WERs.

Perceptions of recognition quality broadly reflected actual performance, but users consistently gave higher quality scores than wizards for the same WER. (See Figure 2). The wizards have direct sight of recognition results, whereas users have only indirect evidence.

WER target	# Wiz	# User	# Task	Completed in time limit	Per-turn WER	Per-dialog WER
None	2	6	24	83 %	0 %	0 %
Low	4	12	48	83 %	32 %	28 %
Med	4	12	48	77 %	46 %	41 %
Hi	2	6	24	42 %	63 %	60 %

Table 3: Summary of experiments

Figure 3 shows results from the wizard understanding status tagging. Misunderstandings increase as WER increases. At the Low and Med levels, wizards are falsely admitting (UnFlaggedMis) utterances at a rate of 5-15%, and rejecting 15-25% of utterances. At the Hi WER level, UnFlaggedMis jumps to 30%, with correct (Full & Partial) interpretations dropping to 35%. Even at the extreme Hi WER condition, wizards are managing to assist users to successfully complete dialogs. We believe this demonstrates the wizard's ability to both parse the ASR output well and assimilate contextual knowledge about what user questions are likely to follow which other questions.

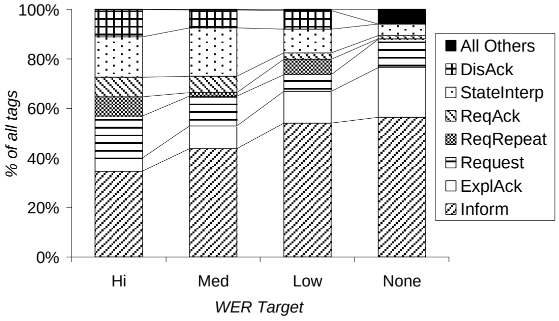

Figure 1: Distribution of wizard actions

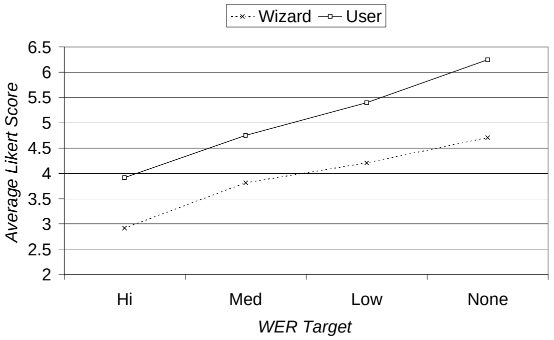

Figure 2: Perception of Recognition Accuracy

4.1. Successful repair patterns

We were interested to see which wizard behaviors were most successful after a dialog problem, akin to [3].

We first separated Wizard turns in which the Wizard was aware of a dialog problem – i.e., FlaggedMis or NonUnderstanding. We next classified the action the wizard chose in these turns into one of 5 strategies, according to Table 4. Finally, as a metric of the outcome of action selection, we determined the wizard understanding status of the *next* wizard turn.

Results are shown in Figure 5. At the Med ASR target, the difference in "Full" alone is significant (*X^2=4.73, df=1, p<0.030):* asking domain-related questions produces more complete, correct understandings. At the Hi ASR target, the difference in UnFlaggedMis is significant (*X^2=4.02, df=1, p<0.045)* and the difference in FlaggedMis+Non (taken as one category) is also significant (*X^2=8.39, df=1, p<0.004).* Asking domain-related questions resulted in more misunderstandings and fewer indicated dialog problems.

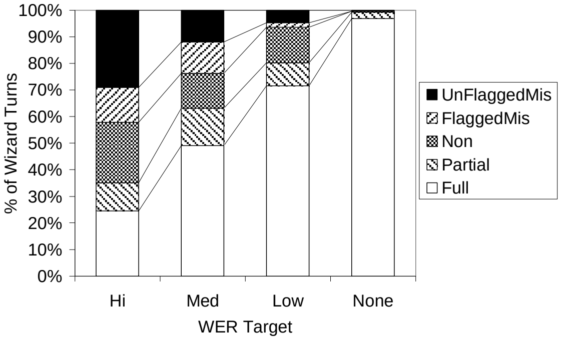

Figure 3: Wizard understanding status

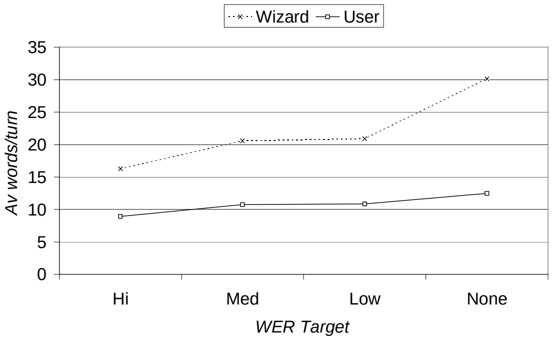

Figure 4: Average turn length (words)

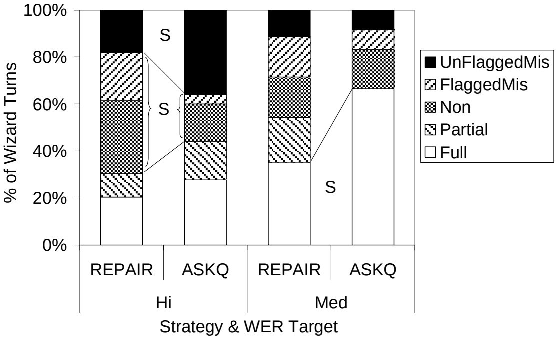

Figure 5: Wizard understanding status one turn after known dialog trouble: effect of 'REPAIR' vs 'ASKQ'. 'S' indicates significant differences discussed in the text.

Label	Meaning	wiz init ?	Tags
REPAIR	Attempt to repair	Yes	*ReqAck, ReqRepeat, StateInterp, DisAck, RejOther*
ASKQ	Ask task-question	Yes	*Request*
GIVEINFO	Provide task info	No	*Inform*
RSPND	Non-initiative taking grounding actions	No	*ExplAck, Rsp-Affirm, Rsp-Negate, Unsolicited-Affirm*
OTHER	Not included in analysis	n/a	*All others*

Table 4: Wizard "Strategies"

Figure 6 omits the NONE and LOW ASR confusion settings as there were insufficient misunderstandings to produce meaningful comparisons. The GIVEINFO strategy is not shown because it was very rarely used following dialog problems. This finding is in contrast to [3], which found that a strategy most akin to GIVEINFO was used frequently (and was most successful). We believe this is a consequence of the differences in the task structure. The tasks in our corpus are a mix of information-seeking and other types, whereas the tasks in [3] are mainly information-giving.

4.2. Characterizing Natural Behavior

We first examined initiative. While intuitively appealing, initiative is a difficult concept to annotate reliably: in [2], the authors found unexpectedly low Kappa scores for initiative tagging. Here we do not attempt to define initiative precisely, nor perform an initiative-specific tagging. Rather, we focus on the wizard, and note that the categories in Table 1: Wizard understanding status categories show evidence of wizard initiative. We regard REPAIR and REQUEST turns as showing initiative, GIVEINFO and RESPOND turns as showing absence of wizard initiative. Results (Figure 6) show a clear trend for the wizard to take more initiative as target WER increases, primarily due to increased grounding activity. However, overall level of wizard control is well below the HC findings in [2], and closer to HH levels.

Another behavior we were interested in is turn length. Figure 4 shows turn length (in words) for wizard and user turns across WER levels. Wizard turn length shortens as WER increases; we believe this is due to increased turns using just grounding acts and fewer providing information. Overall, the ratio of wizard and user words/turn appears closer to HH levels of conversation, based on figures in [2].

A final behavior we were interested in is a user's reaction after an UnFlaggedMis. We separated user turns following an UnFlaggedMis and determined how many included grounding acts signaling trouble – i.e., *RejectOther*, or *DisAck* – and how many included requests for information. Results are shown in Table 5. Surprisingly, users give an explicit signal of dialog trouble in only as many as 20% of turns following an UnFlaggedMis.

4.3. Identifying appropriate reward measures

To identify possible appropriate reward measures, we adopted the PARADISE approach [8]. PARADISE attempts to explain an overall user satisfaction metric using a linear combination of a task completion metric and dialog cost metrics, such as dialog quality and efficiency measures.

We considered several possible task completion metrics. At the end of each dialog, users were asked to what extent

WER target	User turns including tag		
	DisAck	*RejectOther*	*Request*
None	N/A	N/A	N/A
Low	0.0 %	3.8 %	92.3 %
Med	2.5 %	19.0 %	75.9 %
Hi	0.0 %	12.3 %	87.0 %

Table 5: User behavior following an UnFlaggedMis

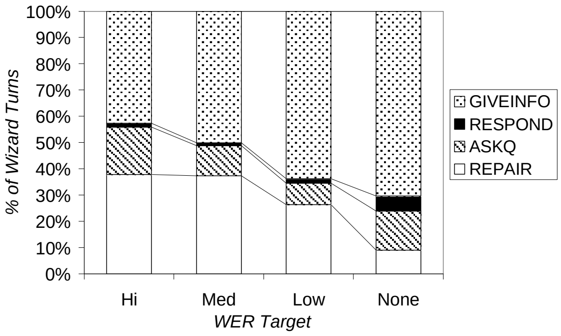

Figure 6: Level of wizard "Initiative"

they believe they accomplished the task. This formed the *User* metric. Later, the tasks were graded on a scale of 1-3 for task precision (accuracy of information) and task recall (coverage). An F-measure of these forms the *Obj* metric.

Experiments of this type suffer from the limitation that users do not receive any evidence whether the information they receive is correct or not. Thus all user feedback may be predicated on a false impression of task completion, which would probably be ultimately discovered in the real world. Further, they may construe the task slightly differently than the experimenter intended. Thus it is not clear to us what measure of task completion is appropriate. We posited a task completion metric to address this issue called *Hyb*. We observe that user's perception of task completion could be interpreted as an indication of how much the user accomplished relative to the task the user *thought* they were supposed to accomplish, forming a measure of recall. *Hyb* is calculated by first normalizing *User* and objective task precision, then combining them in the style of an F-measure.

We considered the following measures of dialog cost: *PerDialogWER, %UnFlaggedMis, %FlaggedMis, %Non, Turns, %REPAIR, %ASKQ.* For overall user satisfaction, we considered two metrics. The *single (S)* metric uses users' responses to a single question eliciting overall satisfaction. The *Combi (C)* metric added together scores from all Likert questions (including perception of task completion).

Table 6 shows results; all R^2 values are significant. In general, the *single* and *combi* metric produced very similar results; where they select the same terms as predictors, we show only the *single* metric. Also, for all regressions (except the "ALL" dataset) which use the *User* task completion metric, *User* was the only significant contribution, and these experiments are not shown. Through regressions examining the *User* and *Obj* metrics (not shown), we found that *user perception* of task completion is a better predictor of overall satisfaction than *actual* task completion, as also noted in [9].

Data set	Metrics (task & user sat)	R^2	Significant predictors
ALL	User-S	52 %	*1.03 Task*
ALL	User-C	60 %	*5.29 Task – 1.54 %UnFlagMis*
ALL	Obj-S	24 %	*-0.49 Turns + 0.38 Task*
ALL	Obj-C	27 %	*-2.43 Turns – 1.45 %UnFlagMis + 1.35 Task*
ALL	Hyb-S	41 %	*0.74 Task – 0.36 Turns*
Hi	Obj-S	40 %	*0.98 Task*
Hi	Hyb-S	48 %	*1.07 Task*
Med	Obj-S	16 %	*-0.62 %Non*
Med	Obj-C	37 %	*-3.35 %Non – 2.94 Turns*
Med	Hyb-S	38 %	*0.97 Task*
Low	Obj-S	28 %	*-0.59 Turns*
Low	Hyb-S	40 %	*-0.49 Turns + 0.40 Task*

Table 6: Select results of PARADISE regressions (see text)

When run on all data, mixtures of *Task, Turns, %UnFlaggedMis* best predict user satisfaction. We believe *%UnFlaggedMis* is serving as a better measurement of understanding accuracy than WER alone, since it effectively combines recognition accuracy with a measure of confidence. These findings are generally consistent with [9]. Broadly speaking, task completion is most important at the High WER level, task completion and dialog quality at the Med WER level, and efficiency at the Low WER level.

Few of the regressions on the None WER target were significant, in part due in part to low levels of variation in the data for the independent parameters under study.

5. Conclusions and Future Work

We have explained the domain, tasks and collection procedures for the SACTI-I corpus. At moderate WER levels, asking task-related questions appears to be more successful than direct dialog repair. Levels of expert "initiative" increase with WER, primarily as a result of grounding behavior. Users infrequently give a direct indication of having been misunderstood, with no clear correlation to WER. Finally, task completion appears to be most predictive of user satisfaction; however, efficiency shows some influence at lower WERs. Future work will apply these insights to statistical systems.

6. Acknowledgements

Funding for the work was provided in part by EU Framework 6 TALK Project (507802). Thanks to Karl Weilheimer and Matt Stuttle for their assistance with the tests and for helpful comments on transcription conventions. Thanks also to Michael Kipp of DFKI for making the ANVIL tool available.

7. References

[1] M. Stuttle, J. Williams, S. Young. (2004). A Framework for Wizard-of-Oz Experiments with a Simulated ASR-Channel. *ICSLP. Jeju, South Korea.*

[2] C. Doran et al. (2001). Comparing Several Aspects of Human-Computer and Human-Human Dialogues. *Proc. 2nd SIGDial Workshop.* Aalborg, Denmark.

[3] G. Skantze. (2003). Exploring Human Error Strategies: Implications for Spoken Dialogue Systems. *ISCA Workshop on Error Handling in Spoken Dialogue Sys.*

[4] Linguistic Data Corporation. (2003). Guidelines for RT-03 Transcription. On-line Manuscript, version 2.2.

[5] K. Maeda et al. (2002). Creating Annotation Tools with the Annotation Graph Toolkit. *Proc 3rd Intl Conf on Language Resources and Evaluation, pp 1914-1921.*

[6] M. Kipp. (2001). ANVIL – A Generic Annotation Tool for Multimodal Dialogue. *Proc. Eurospeech*.

[7] D. Traum. (1994). *A Computational Theory of Grounding in Natural Language Conversation.* Ph D Thesis, Dept of Computer Science, Univ of Rochester.

[8] M. A. Walker et al. (1997). PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *Proc ACL/EACL, San Francisco, pp. 271-280.*

[9] M. A. Walker et al. (1998). Towards Developing General Models of Usability with PARADISE. *Natural Language Engineering 1(1), pp 1-16.*