A WOz Variant with Contrastive Conditions[1]

Esther Levin[†] *and Rebecca Passonneau*[‡]

[†]Department of Computer Science, City College of New York, USA
esther@cssfs0.engr.ccny.cuny.edu
[‡]Center for Computational Learning Systems, Columbia University, USA
becky@cs.columbia.edu

Abstract

We present a variant of the WOz paradigm we refer to as incremental ablation. The new feature involves incrementally restricting the human wizard's capacities in the direction of a dialog system. We lay out a data collection design with six conditions of user-system and user-wizard interactions that allows us to more precisely identify how to close the communication gap between humans and systems. We describe the application of the method to analysis of contexts in which ASR errors occur, giving us a means to investigate the problem-solving strategies humans would resort to if their communication channel were restricted to be more like the machine's. We describe how we can use the methodology to collect data that is more relevant to a particular learning paradigm involving Markov Decision Processes (MDP).

1. Introduction

With more than 8000 applications worldwide, the market for automated dialog systems continues to grow [1]. However, the chasm between research and industry dialog systems is arguably as large as the difference between human-human and human-machine dialog. In addition, callers often fail to complete such dialogs, despite the fact that word error rate (WER) for ASR has continued to drop over the past decade. The recent focus on reducing WER coincides with the rise of hidden Markov modeling and an increasing emphasis on data-driven approaches to language modeling [2, 3]. However, without corresponding improvements in other parts of dialog systems, the payoff in system performance will not be commensurate with the effort expended in reducing WER.

By analogy to [2], which asks what we can learn from human speech perception, we believe it would be fruitful to look more closely at human action in dialog. To do so, we need a more fine-grained methodology than currently exists for analyzing dialog data, one that would make it possible to isolate different aspects of behavior, such as the analysis of the repertoire of dialog actions as distinct from the mappings between actions and utterance form, strategies for combining or merging actions, and of contextual features that condition actions and utterance forms. In this paper, we present a novel paradigm for investigating dialog strategies in the presence of ASR errors that provides such a fine-grained methodology.

Our general goal is to provide a means to investigate ASR errors by asking what dialog acts and what dialog strategies would be more likely to help advance the dialog in a manner that humans would find sufficiently natural. Our specific goal is to have a means to investigate the problem-solving strategies humans would resort to if their communication channel were restricted to be more like the machine's, or if there were other limitations to their communicative functionality that were closer to the machine's limitations. We propose a variant of WOz in which we incrementally modify the wizard in the direction of the machine in order to compare the wizard's behavior under a variety of conditions. We refer to our paradigm as incremental wizard ablation. It draws on two methods in AI that have a long history, ablation and comparison studies [4], and integrates them within the WOz paradigm, in order to isolate different aspects of human-machine interaction for study. It resembles an ablation study, in which a system's performance with and without a given component is contrasted in order to study the contribution of the component's functionality to overall performance [4]. But instead of removing a component from a system, we remove some dimension of the wizard's communication resources and replace it with the corresponding resource used by the system. It also resembles a comparison study [4], in which a given system functionality is approached in one or more ways, e.g., comparing several parsers against each other to compare their performance within a system, or on different corpora. But instead of comparing a system component that has been engineered using alternative methods, we will compare how a system versus a human uses a given resource.

In this paper, we describe how we can use the methodology to collect data that is more relevant to a particular learning paradigm involving Markov Decision Processes (MDP). Previous work in this paradigm has shown that dialog strategies can be learned from data for a given set of dialog actions, and a given representation of the dialog state [5-7] In principle, our method will make it possible to investigate the state representation, the dialog acts and the strategies somewhat independently of each other.

We identify three benefits of incremental wizard ablation:

1) Currently, the user models for MDP or POMDP (Partially Observable MDP) approaches are typically estimated from dialog corpora that are treated as a single exemplar that captures the range of expected behaviors. Our methodology will result in contrastive corpora that differ in controlled ways.

[1] Author order is alphabetical.

2) WOz studies have been widely used in developing spoken language systems [8] but primarily at the early stages. By integrating WOz studies into the system development cycle, we can predict more precisely what system enhancements should produce the greatest performance gain

3) Methods for analyzing dialog act types and dialog strategies do not address the differences between human communicative capacities and machine communicative capacities. While this remains a distant target, we believe our methodology will yield new insights into the potential for human-machine interaction.

Our paper is organized as follows: section 2 discusses related work in which wizards have been restricted. The findings strongly support the view that systems will become more habitable if they focus less on individual recognition errors and more on the dialog task. Section 3 reviews the schematic architecture of a dialog system. We have identified a real world application involving collaboration with a library whose patrons do most of their borrowing over the phone. Section 4 gives a brief summary of the domain our experiments will be conducted in, and the application we aim to build, an Automated Readers Adviser (ARA). Section 5 presents the incremental wizard ablation model in detail, and lays out the design of our experiments. In section 6 we discuss how the experiments can enrich an MDP approach.

2. WOz methodology and its extensions

The original goal of WOZ data collection was to learn how users interact with an intelligent 'automated' system; however, dialogs collected through such WOZ studies rarely exhibit problems that are typical to human-machine interactions, such as misrecognition or misunderstanding of the sort machines make. Recently the notion of ablation was introduced in WOZ studies by extending the WOZ paradigm to study both user and wizard behavior in face of such misrecognition and misunderstanding errors.

In [9,10], a study of error recovery strategies in human-human dialogs was performed, where, in order to elicit error handling strategies, a speech recognizer was used to process the speech of the user, and the wizard could read the recognition results, but not hear the utterances. The results indicate that speech recognition errors caused relatively few misunderstandings, but many non-understandings. Following [9] *misunderstanding* means that one participant mistakenly believes that she has a correct interpretation of the other participant intention. When she fails to obtain any interpretation at all, or obtains more than one interpretation with no way to choose among them, a *non-understanding* has occurred. One important difference between non-understandings and misunderstandings is that non-understandings are recognized immediately by the addressee, while misunderstandings may not be identified until a later stage in the dialogue or not detected at all. The low incidence of misunderstandings in [9, 10] suggests that different knowledge sources (such as confidence estimations, syntactic structure and context) can be used (at least by humans) for detection of errors in the speech recognition result, and for deciding upon appropriate reactions to them. In addition, it was found that for the task of navigation and map directions the human operators' most effective error handling strategy was to ask task-related questions instead of signaling non-understanding (See Table 1 for examples).

In [11] WOZ studies were conducted for a tourist information domain. Again, the wizard could not hear the user, but had access to the user's utterance processed by a simulated ASR channel with controlled and varying word error rate [12]. The researchers found that even at the maximum word error rate condition, wizards managed to assist users to successfully complete dialogs. The results of the study also showed that for this domain the most successful error-handling strategy was asking task-related questions, rather than engaging in explicit error sub-dialogues. Similar finding comparing different strategies in situation involving non-understanding were recently reported in [13].

These studies suggest that dialog systems designed to imitate successful error handling strategies of human wizards by focusing on the task rather than signaling non-understanding will enjoy higher task success rate and user satisfaction. We generalize the notion of wizard ablation in two ways: first, the ablation is performed incrementally to better isolate the different aspects of dialog management we are trying to learn; and second, data collection is performed in different phases of the system development cycle thus integrating data collection, development and evaluation.

Strategy 1: Signal of non-understanding	U: **west with** *(That's right.)* O: Please repeat what you said.
Strategy 2: Task-related question about position	*O: Do you see a wooden house in front of you?* *U:* ***yes crossing address now*** *(I pass the wooden house now.)* *O: Can you see a restaurant sign?*

Table 1: The different operator strategies after non-understandings (from [9]).

3. Spoken dialog system functional components

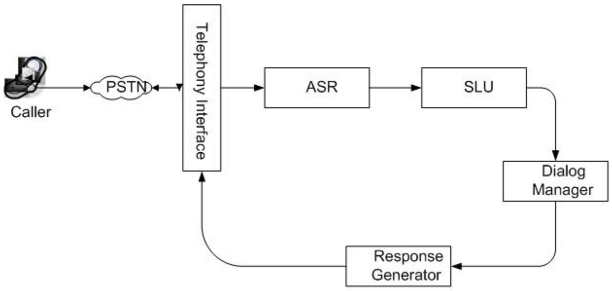

Figure1 : Functional Block Diagram of Dialog System

Figure 1 illustrates a generic functional block diagram for a dialog system. When the user speaks, the signal captured by the telephony interface is passed through the Automated Speech Recognition (ASR) module that converts the waveform representing user utterance to text hypotheses. The recognition result is passed to the Spoken Language Understanding (SLU) module that translates it to some formal representation of the meaning for this utterance. The state of the dialog is maintained by the Dialog Manager (DM) that updates it with each new user input obtained from SLU and decides according to its dialog strategy on one of finite set of preprogrammed dialog actions to move the dialog forward. The Response Generator converts this action first to verbalized text and then to voice prompt the user hears concluding the current dialog exchange and possibly starting the next one.

4. Automated Readers Adviser (ARA)

As noted in section 2, task oriented dialogs can be successful without resolving the interpretation of every utterance. We have identified a real-world application domain with very highly motivated users. Our goal is to automate a subset of the calls patrons make to the Andrew Heiskell Braille and Talking Book Library in New York City. It provides books in alternative formats, such as audiotape and Braille, for patrons with various visual or physical impairments.

The same issues that make these patrons eligible for borrowing privileges at Heiskell makes it difficult for them to use the library in person, either because travel is more difficult than for the average person, or because they cannot browse library materials unassisted. Much of their interaction with the library is by telephone, including most of the borrowing transactions. We hypothesize that an Automated Readers Adviser (ARA) could handle routine requests (such as finding a book by title or by author and perhaps reserving it), more quickly than human librarians, because database access would be immediate. This would free the patrons and librarians to use their phone interactions for more complex requests.

5. Incremental Wizard Ablation

Six conditions of data collection in the incremental wizard ablation paradigm are presented in Table 2. These conditions are designed to provide data for a particular learning paradigm involving MDP. However, the general approach could be applied by ablating other wizard functions, to learn other aspects of dialog behavior. Under all conditions specified in Table 2 overall dialog performance will be measured using task success (ordering a library item or set of items), time on task, and user satisfaction, as in the Paradise model [14].

DM Status	User-AW1 Wizard	User-AW2 Wizard	User-System
I:Baseline	A	B	C
II:Enriched	D	E	F

Table 2. Schematic table of six dialogue conditions, illustrating dimensions of contrast

We will have two phases of data collection using scenarios from the Automated Readers Adviser domain. Each phase will contrast the system performance in ordinary interactions with users (column 4 User-System) with the performance of ablated wizards (AW) under two ablation conditions (columns 2 and 3).

We can define the three columns of Table 2 with respect to the functional block diagram of a dialog system shown in Figure 1. Under standard WOz conditions, the wizard replaces all dialog system components except the response generator. The Wizard uses his or her human intelligence to perform the functionality of speech recognition, spoken language understanding and dialog management, maintaining as much knowledge of the dialog history in his/her state as necessary, and using dialog actions he/she feels best for the situation. In AW1, in addition to being constrained to use the system's response generation, the wizard will be restricted to the same inputs from the ASR and SLU components that the system receives. The wizard's auditory channel is ablated and replaced with the corresponding system module. The AW1 wizard will otherwise have freedom to interact as a standard wizard. In AW2, the wizard will be additionally restricted to the same set of Dialogue Actions (DAs) available to the system's Dialog Manager (DM). Here both the wizard's auditory channel and action set are ablated and replaced with the corresponding system modules. By comparing dialogs across columns, we can isolate the effects of different aspects of the communication devices (wizard versus system) on the success of the dialog, as well as on the distribution of ASR errors and consequent dialog acts. By comparing dialogs across rows, we can isolate the effects of system enhancements on three types of interactions.

The results of Phase I data collection, directed at hypothesis generation, will be used to modify the Phase II DM (used by both system and AW2 Wizard), and if it seems required, to add new scenarios. The Phase II data collection will include testing of specific hypotheses yielded during Phase I, as well as continued hypothesis generation and refinement. We will then experiment with different subsets of dialogs from each condition to investigate the impact on automated learning of dialog strategies in the MDP framework.

The conditions in Table 2 contrast with respect to whether the wizard freely creates responses or composes them from DM choices (cells A/D versus B/E); whether actions are constrained, independent of whether the "expert" is a wizard or the system (cells A/D versus B/C/E/F); whether the DM (used by System and AW2 Wizard) has been enriched (cells B/C versus E/F); and by human wizard versus system dialogue strategy(cells B/E versus C/F)

We hypothesize that *incremental wizard ablation* will allow us to more precisely evaluate where the biggest differences are between wizards and systems, and discover avenues for narrowing the gap. The three areas where we will focus our investigation will be on identifying a need for a different repertoire of Dialogue Acts (DAs), a more flexible means of selecting and combining DAs, or a better representation of open goals in the context (e.g., use of a stack to simulate certain aspects of planning). However, the exact locus of our efforts will depend on the nature of the results we find in Phase I.

Consider the case where our results show that the average performance difference between conditions B and C is much greater than that between A and B. This would suggest that the system could do better without changing the basic components of the DA set, since the AW2 wizard in condition B was able to perform well with the same DAs as the system in condition C. We would then need to ask questions of the following sort:

1. Did the AW2 wizard make different DA choices than the system? We would address this question by performing a distributional analysis of types and frequencies of DAs in dialogues from conditions B versus C.
2. Did the AW2 wizard use specific aspects of context in selecting DAs that are not available to the system? We would address this question by investigating what aspects of context seem to determine AW2's behavior.

On the other hand, consider the case where the average performance difference between cells A and B is much greater than that between B and C. This would suggest that the set of DAs in the DM is too impoverished even for a wizard, and our efforts would be focused on a contrastive analysis of dialogues from conditions A and B to determine how to enrich or otherwise modify the types of DAs to bring them closer to those that the AW1 wizard uses.

To measure improvements in system performance due to changes in the DM (conditions F versus C), we can ask: Do user-system dialogues in Phase II compared with those in Phase I have a higher rate of task success? A reduction in time on task? An increase in user satisfaction?

To see whether the enhanced DM in Phase II has narrowed the gap between AW2 wizards (B/E) and system (F/C), we can ask, do user-AW2 dialogues compared with user-system dialogues have the same or different changes (Delta E-B versus Delta F-C) in rate of task success? in time on task? in user satisfaction?

In sum, by the end of Phase II, data for six conditions will be collected. Our paradigm will allow us for the first time to generate more focused hypotheses about how well current dialogue systems could perform if they retained the same input (ASR/SLU) and output channels (response generator), but were able to learn in more focused way from human dialogue actions and strategies.

6. Learning from Data

Markov Decision Models (MDP) [16] and Partially observable MDP models have been used to model dialogue in terms of its *action set*, *state space*, and *dialogue strategy.* The action set of the dialogue system includes all possible actions it can perform, such as interactions with the user, interactions with external resources (e.g. querying a database), and internal processing. The state *s* of a dialogue system includes the values of all the relevant internal variables that determine the next system action. Given the same external conditions (i.e. user responses, database results, etc.) the next system behavior is uniquely determined by the current state. The dialogue strategy specifies, for each state reached, the next action to be invoked by the system.

Although it was shown [6- 8] that within this framework it is possible for a given action set and state representation to learn dialogue strategy from data in order to improve the overall system performance, little effort has been dedicated to research methodologies for defining the *right* state space and action set. By limiting the wizard's input to the same one the DM has access to (the output of ASR and SLU) we can isolate the dialogue level features the wizards are taking advantage of in their decision making and use these features to enhance the state representation of DM. By contrasting the conditions AW1 and AW2, of limiting the wizard to a finite set of dialogue actions and allowing him to freely use any action, we can study how to enhance the action set of the DM in order to mimic the wizard performance.

7. Conclusion

Two key observations motivate wizard ablation. First, the types of errors that occur in human-human dialogue will not be representative of human-machine dialogue. Second, a wizard study is already an interaction involving a human wizard who has been ablated in the direction of a dialogue system: the wizard uses a dialogue system's response generator. By incrementally ablating the wizard to rely increasingly on system components rather than human capabilities, we will have the opportunity to learn from the types of errors and misunderstandings that occur, and the impact they have on the success of the dialogue. More importantly, we can control for a wider range of contrasts in the interactions we observe, and begin to study the important question of how to create datasets specifically for the purpose of learning dialogue strategy.

8. References

[1] R. Pieraccini and J. Huerta, "Where do we go from here? Research and Commercial Spoken Dialog Systems", in Proc. SIGDIAL. Lisbon, Portugal, September 2005.

[2] S. Dusan, L.R. Rabiner, "Can automatic speech recognition learn more from human speech perception?" In C. Burileanu, ed., *Trends in Speech Technology*, pp. 21-36. Romanian Academy Publisher, Bucharest, Romania., 2005.

[3]R.K. Moore, "A comparison of the data requirements of automatic speech recognition systems and human listeners. Eurospeech 03, pp. 2582-2584. Geneva, 1-4 September 2003.

[4] P.Cohen, and A. Howe. How evaluation guides AI research. AI Magazine, winter 1988, pp. 35-43., 1988.

[5] E. Levin and R. Pieraccini, "A stochastic model of computer human interaction for learning dialogue strategies," in *Proc Eurospeech*, Rhodes, Greece, 1997, pp. 1883–1886

[6] K Scheffler and SJ Young, "Corpus-based dialogue simulation for automatic strategy learning and evaluation," in *Proc NAACL-2001 Workshop on Adaptation in Dialogue Systems*, Pittsburgh, USA, 2001.

[7] S Singh, DJ Litman, M Kearns, and M Walker, "Optimizing dialogue management with reinforcement learning: Experiments with the njfun system," *Journal of Artificial Intelligence Research*, 2002.

[8] N.Fraser, and G.Gilbert, "Simulating speech systems." Computer Speech & Language 5: 81-99, 1991.

[9] G. Skantze. "Exploring human error handling strategies: Implications for spoken dialogue systems." In *Error Handling in Spoken Language Dialogue Systems*. International Speech Communication Association, 2003.

[10] Skantze, G. "Exploring Human Error Recovery Strategies: Implications for Spoken Dialogue Systems." *Speech Communication*, 45(3) (pp. 207-359), 2005.

[11] J. D. Williams and S. Young. "Characterizing Task-Oriented Dialog using a simulated ASR Channel. " ICSLP, October 2004, Jeju, South Korea.

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

[13] D. Bohus and A.I. Rudnicky, "Sorry, I Didn't Catch That! - An Investigation of Non-understanding Errors and Recovery Strategies", in Proc. SIGDIAL. Lisbon, Portugal, Sep. 2005.

[14] M. Walker, D.J. Litman, C. A. Kamm, A. Abella, "PARADISE: A Framework for Evaluating Spoken Dialogue Agents", In Proc. of the 35th Annual Meeting of the Association of Computational Linguistics, ACL, 1997.

[15] S. J. Young, "Talking to Machines (Statistically Speaking)", in Proc. ICSLP-2002, Denver, Colorado, 2002.

[16] Levin, E., Pieraccini, R., Eckert, W., “A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies, “, IEEE Trans. on Speech and Audio Processing, v. 8, n.1, 2000