The Influence of User Tailoring and Cognitive Load on User Performance in Spoken Dialogue Systems

Andi Winterboer[1]*, Jiang Hu*[2]*, Johanna D. Moore*[1]*, Clifford Nass*[2]

[1]School of Informatics, University of Edinburgh, United Kingdom
[2]Communication between Humans and Interactive Media Lab, Stanford University, CA
A.Winterboer@ed.ac.uk, Huj@stanford.edu, J.Moore@ed.ac.uk, Nass@stanford.edu

Abstract

This paper presents results of a Wizard-of-Oz (WoZ) study examining the effect of two different information presentation methods on a secondary task, namely driving. The results demonstrate that the user-model based summarize and refine (UMSR) approach enables more efficient information retrieval in comparison to the data-driven summarize and refine (SR) approach, and does not negatively affect driving performance.
Index Terms: spoken dialogue systems, natural language generation, content selection, information presentation

1. Introduction

Spoken dialogue systems (SDS) are developed to provide natural, efficient and user-friendly access to applications and services (e.g., calendars, public transport information, flight booking, product recommendation) using speech as the main interaction mode. In contrast to graphical user interfaces, SDS challenge human memory and attention since they present information sequentially and non-persistently. Because many SDS are intended for use in situations where the user is performing another task simultaneously, e.g., driving a car, it is crucial that intelligent algorithms for effective information presentation are devised. In the context of in-car information systems, where safety is of paramount importance, distraction effects must be minimized in order to guarantee that driving behaviour is not adversely affected. Therefore, an assessment of the cognitive load imposed by different information presentation strategies is a crucial factor in the design and development of in-car voice services. To explore this question, we performed a WoZ experiment comparing two previously proposed approaches to information presentation, focusing on their effect on driving-related cognitive load.

The summarize and refine (SR) approach to information presentation, developed by [1] and later extended by [2], groups a large number of options into a small number of clusters that share attributes. Then, the system summarizes the clusters based on their attributes and suggests additional constraints to the user. For large data sets, attributes that partition the data into the minimal number of clusters are chosen, so that a concise summary can be presented to the user to refine. However, as argued in [5] there are several limitations to this approach. First, many turns may be required during the refinement process. Second, if there is no optimal solution, exploration of trade-offs is difficult. Finally, the attributes on which the data has been clustered may be irrelevant for the specific user. Table 1 shows an example dialogue based on the SR approach.

The user-model based (UM) approach employs a user model and decision theory techniques to identify and present a small number of options that best match the user's preferences [3, 4]. This approach to information presentation can concisely present a relatively small number of options, highlighting the ways in which those options satisfy user preferences. However, it does not scale up to presenting a large number of options. In addition, users may not be able to provide constraints until they hear more information about the space of options. Finally, because this approach does not provide an overview of the available options, it may lead to the user's actual or perceived missing out on potentially better alternatives.

Table 1: *SR dialogue sample*

User:	"I would like to go from San Francisco to Prague on January 30^{th}."
SR:	"I found 21 flights from San Francisco to Prague. All these flights require a connection. There are flights available in economy, business, and first class. I also have information about price range."
User:	"I'd like a business class flight."
SR:	"I found 11 business class flights from San Francisco to Prague. All these flights require a connection. There are flights in the expensive and moderate price range. I also have information about arrival times."
User:	"I'd like to arrive before 3 p.m."
SR:	"..."

The user-model based summarize and refine (UMSR) approach [5] was developed to combine the benefits of the UM and SR approaches, by integrating user modelling with automated clustering. When there are more than a small number of relevant options, the UMSR approach builds a cluster-based tree structure which orders the options to allow for stepwise refinement. The effectiveness of the tree structure, which directs the dialogue flow, is optimized by taking the user's preferences into account. Trade-offs between alternative options are presented explicitly to give the user a better overview of the option space and lead the user to a more informed choice. To give the user confidence that they are being presented with all relevant options, a brief account of the remaining (irrelevant) options is also provided.

Results of a laboratory experiment comparing the UMSR and SR presentation approaches demonstrated that participants clearly preferred the UMSR approach [5]. In this study, participants were asked to read and evaluate transcripts of six dialogue pairs. Each pair consisted of one dialogue using the SR approach to information presentation and one using the UMSR

approach. The four criteria used for evaluation were:

- **understandability** ("Did the system give the information in a way that was easy to understand?"),
- **overview of options** ("Did the system give you a good overview of the available options?"),
- **relevance of options** ("Do you think there may be flights that are better options for the user that the system did not tell her about?"), and
- **efficiency** ("How quickly did the system allow the user to find the optimal flight?").

Although the presentations based on UMSR were found to be favored, it remained an open question whether such preferences would still be observed when the user is actually conversing with an SDS, especially when conducting another task, such as driving, at the same time.

To shed light on this question we conducted a Wizard-of-Oz experiment to compare these approaches in situations of low vs. high workload with a simulated SDS [6]. A total of four driving courses with two levels of difficulty were used to vary driving-related cognitive load. Compared to the easy courses, the difficult courses had three times as many vehicles, cyclists, and pedestrians as well as sharp curves, two foggy sections, a construction site, slopes of various degrees, and a police chase. It was assumed that the basic principles of the UMSR approach would lead to more efficient dialogues, fewer harmful distractions to drivers and a pleasant user experience in comparison with an SR-based SDS. However, the results, based on data of driving safety, evaluation of the spoken dialogue system, and perception of self, were mixed and only partially supported these hypotheses. In terms of dialogue efficiency, the UMSR-based system was superior to the SR-based one. The average dialogue duration was significantly shorter. Additionally, participants in general required fewer dialogue turns when the system adopted the UMSR approach than when it utilized SR. In contrast, in terms of driving safety, participants had significantly more minor errors when the system adopted the UMSR approach than when it used SR. Finally, concerning the perception of the system and self, the obtained results were twofold. For participants driving on easy courses, UMSR was perceived as "more fun". When participants drove on the difficult courses, UMSR led to more driving errors and SR was preferred. Thus we see that the cognitive load involved in the driving task, most notably when participants drove on a difficult course, affected the user's perception of the SDS when it utilized UMSR-based presentations. After thoroughly examining the transcribed dialogues we discovered a possibly critical flaw in the implementation of the deployed UMSR algorithm. Some first-turn UMSR presentations were very long, making them difficult to comprehend, and also contained redundant information which placed a large and unnecessary cognitive burden on the participants, particularly when driving-related cognitive load was already high. Thus, we implemented a new version of UMSR in which we controlled for information density and turn length in order to perform a fair comparison with the SR approach.

2. Experiment

Due to the observed problems caused by the long system turns, a newly implemented variant of UMSR, presented in this paper, was introduced. Tackling the driving safety issue as well as improving users' perceptions of the system and self were the main motivations for conducting this follow-up study. Table 2 shows an example dialogue employing the refined UMSR approach.

Table 2: *UMSR dialogue sample*

User:	"I'd like to go from San Francisco to Prague on January 30^{th}, please."
UMSR:	"There are no direct flights from San Francisco to Prague, but I found 11 flights with availability in business class. 2 of these are on KLM."
User:	"Do you've anything arriving before 3 p.m.?"
UMSR:	"There are 2 business class flights from San Francisco to Prague on KLM that will get you there on time. The first flight arrives at 2 p.m. with a total travel time of 18 hours and 25 minutes. The second flight arrives at 2:45 p.m. with a total travel time of only 14 hours."
User:	"I'd like the second flight, please."

Since the UMSR approach was successful for drivers of the easy courses (in terms of task success and partly in terms of user satisfaction), participants were asked to drive exclusively on the difficult courses for this study.

2.1. Participants

A total of 16 students from Stanford University, all licensed drivers, were paid to participate in the experiment. Participants with prior exposure to the driving simulator were excluded; gender was balanced across conditions.

2.2. Experimental setup

The STISIM Drive^{TM} simulation system with projected visuals on a wall-sized back-projection screen was deployed to simulate as realistic as possible driving courses. To keep track of each participant's driving performance, numbers of collisions, speeding tickets, stop sign violations, and minor driving errors (i.e., centerline crossing and road edge excursion) were recorded. The two difficult driving courses used in [6] were re-used, containing four sequential sections: A residential area, a small city, a country highway, and a big city.

2.3. Wizard environment

The Wizard-of-Oz approach [7] provides the opportunity to test hypotheses about not yet implemented systems, such as complex spoken dialogue systems, by simulating the system. In this study, a database-driven Web interface was deployed which automatically generated system responses based on either the SR or the UMSR strategy to presenting information. The wizard was used to perform speech recognition and natural language understanding. Furthermore, the wizard kept the dialogue going if the user was silent. The integrated database contained actual flight information as provided by airlines. The wizard used drop-down menus to perform stepwise queries according to participants' requests until a satisfying flight was found and booked. In the UMSR condition, in order to present information based on a user model, users were asked to role play and were given a business traveler's persona (described below). Textual information provided by the Web interface was copied-and-pasted by the wizard to Speechify^{TM}, a text-to-speech application provided by Nuance Communications, Inc. All participants heard a synthetic voice of their own gender. They were encouraged to talk naturally rather than merely responding to

system prompts. Hence, the wizard used very few questions as prompts and would add additional questions only if the participant remained silent for more than five seconds after each round of information presentation by the system.

2.4. Procedure

Each participant drove for two experimental rounds and booked four different one-way flights. Before the actual experiment, to enable reliable and rigorous comparisons, all participants were briefed to act as a business traveler for the flight booking task. In descending order of importance, the business traveler 1) prefers flying *business class*, 2) is concerned about *arrival time*, *travel time*, and *number of stops*, and 3) wants to fly on *KLM* if possible. In addition, the participants received detailed instructions regarding the two flights to be booked prior to each round of driving. To make the booking process more realistic, the four routes (i.e., pairs of cities) were carefully chosen in order to guarantee that each participant experienced four different scenarios: 1) no KLM flight was available, 2) one KLM flight matched all the criteria, 3) one KLM flight in business class was available but required a connection, and 4) one KLM flight was found but it was in economy class. The order in which the four flights were booked was rotated to counter-balance possible order effects. The order of each participant's two courses was also randomized. In the first round of driving, half of the participants obtained flight information presented from the system adopting the SR approach; the other half received search results presented with the UMSR approach. The opposite approach was used during the second round of experimental driving. Before the experimental phase, participants took a test drive on a demo course to familiarize themselves with the simulator. The following experimental phase consisted of three major steps. In Step 1, the participant was informed that she would interact with an 'in-car information system to book flights while driving. She was instructed to pretend that she was "a business traveler" and then learned about the details of the persona. At the same time, she received instructions on booking the first two flights, including a short story explaining the business traveler's motivation to travel to the specific destination.

In the second step, the participant drove on the first experimental course alone in the lab; the wizard was sitting in a neighboring room. Approximately three minutes later, a short beep was played, followed by the first system utterance saying that "This is the in-car information system. I'm now connected to the network. Would you like to book a flight?" A conversation began as soon as the participant responded to this prompt. Via wireless connections, the wizard monitored all audio events around the driving simulator, performed database queries, and converted textual output into synthetic speech on a laptop computer. The synthetic speech utterances were transmitted to a pair of speakers next to the simulator. After confirming the booking of the first flight, the participant was prompted and continued to book the second flight.

In Step 3, the experimenter returned to the lab and asked the participant to complete a questionnaire evaluating the "in-car information system", the driving course, and the participants' perception of themselves during the interaction with the SDS. After that procedure, Steps 1 through 3 were repeated, with different flights to book, and a different course (of the same degree of difficulty) to drive, and a different presentation method, i.e., SR participants in Round 1 used UMSR in Round 2, and vice versa. After completing the last questionnaire, the participant was debriefed, paid, thanked, and discharged.

3. Results

Dialogues were recorded, transcribed, and analyzed. Data captured by the driving simulator and the questionnaires were tabulated and analyzed in SPSS. For the questionnaire data, ten-point Likert scales were used except for the four seven-point Likert scales used in [5]. The ten-point scales were meant to capture subtle variations and to avoid a middle point that often encourages "satisficing" [8].

3.1. Dialogue efficiency

The mean number of turns each participant required for booking a flight with the system adopting the UMSR strategy (as shown in Table 3) remained relatively unaffected by the conducted modifications. The slight increase in number of turns can be explained by the shorter turn length which necessarily resulted in a higher number of required turns. Still, participants using UMSR took significantly fewer turns than when using the SR-based search system ($p < .05$, indicated with a "*" below).

Table 3: *Number of turns per booking in difficult driving condition*

	Experiment 1 (N=16)	Current experiment (N=16)
SR	16.44*	16.06*
UMSR	11.80*	12.94*

Interestingly, average dialogue duration for SR as well as for UMSR are reduced in comparison with the first WoZ experiment (see Table 4). Again, the significant difference between duration of UMSR (323 seconds) and SR (423 seconds) remains roughly the same ($p < .05$).

Table 4: *Average dialogue duration for 2 bookings in difficult driving condition*

	Experiment 1 (N=16)	Current experiment (N=16)
SR	457 secs*	423 secs*
UMSR	379 secs*	323 secs*

In addition to dialogue duration as a measure of task success, we also counted how often the flight "best" matching the business traveler's profile was chosen. The results are shown in Table 5. Of the 32 flights that were booked with the SR-based system, the most suitable flight was booked in only approximately 60% of the cases. In comparison, the participants booked the most suitable flight in circa 80% of the cases with the UMSR-based system.

Table 5: *How often was the "best" flight selected?*

	Experiment 1 (N=16)	Current Experiment (N=16)
SR	21 (32) (65.625%)	19 (32) (59.375%)
UMSR	23 (32) (71.875%)	26 (32) (81.25%)

In sum, the average flight booking process with a system based on UMSR had a considerably shorter dialogue duration and required fewer dialogue turns. Moreover, in more cases

the best available flight was selected. Thus, information access with the UMSR approach is more efficient than with the SR approach.

3.2. Driving safety

Whereas we found in the first WoZ experiment that participants had significantly more minor driving errors when the system adopted the UMSR approach than when it utilized the SR approach (even though this difference was mainly due to the difference observed among easy-driving participants), this time there were no observable differences between the driving performance of participants in the two conditions in terms of numbers of collisions, speeding tickets, traffic light or stop sign violations, or minor driving errors.

3.3. Perception of system and self

In the obtained questionnaire data we found no significant differences between UMSR or SR concerning the participant's perception of the system, the driving course or self. Answers to the four questions (concerning understandability, overview of options, relevance of options, and efficiency) used in the previous study [5] were also analyzed. All four questions concerning the UMSR presentations received higher average scores than they did in the first experiment. Nevertheless, no significant difference between the UMSR-based and the SR-based system was found.

4. Discussion

The results of the previously conducted studies, asking participants to evaluate presentations based on SR and UMSR presented as dialogue transcripts [5] or as sound files where the participants "overhear" the dialogue [Moore, personal communication] demonstrated a clear preference for UMSR. In the current study, no significant differences on the four user satisfaction questions were found. However, these questions were asked at the end of a list of 85 evaluation questions about the participants' perception of the in-car system, the driving course, and themselves. The sheer number of questions may have affected participants motivation for answering them accurately. In addition, in contrast to the previous studies, participants in this experiment were actively interacting with the spoken dialogue system while conducting another very demanding task simultaneously. In such conditions, participants may be more concerned with completing both tasks, and less able to make subtle distinctions between systems. However, with the refined UMSR approach, there were no significant differences in the number of driving errors between UMSR and SR. This shows that in prior experiments the confounding factor was the length of the UMSR presentations (rather than the user-model controlling the choice of attributes) making it difficult for the participants to comprehend the presentations, especially in unfavorable driving conditions involving high cognitive workload. Therefore, it was necessary to run the follow-up study with a modified UMSR algorithm controlling for turn length and information density. In addition, dialogue duration was significantly shorter with the refined UMSR approach, and users were more likely to pick the best option. Thus we see that the refined UMSR approach is equivalent to SR in terms of user satisfaction and driving safety, but better in terms of task success and dialogue duration.

5. Conclusion

We presented the results of a WoZ experiment comparing two different approaches to information presentation in spoken dialogue systems. In line with results from previous experiments we found that in terms of task efficiency the user-modeled summarize and refine (UMSR) approach clearly outperforms the summarize and refine (SR) approach and enables more effective information retrieval. In contrast to previous experiments where participants focused solely on the flight booking task [5], we have shown that this finding also applies to situations in which another highly demanding task is conducted simultaneously.

In our dual task experiment, we did not see the significant preferences for UMSR that were obtained in prior studies in which participants read or overheard dialogues. To determine whether this is due to the fact that participants were actually interacting with a SDS, or whether it is because they are interacting with the SDS while performing a demanding secondary task, we will conduct an experiment in which participants only interact with the simulated SDS. In addition, we are integrating findings from psycholinguistics to make the generated spoken messages easier to process and comprehend.

6. References

[1] J. Polifroni, G. Chung and S. Seneff, "Towards the Automatic Generation of Mixed-Initiative Dialogue Systems from Web Content", in Proceedings of Eurospeech '03, Geneva, Switzerland, September 2003.

[2] G. Chung, "Developing A Flexible Spoken Dialog System Using Simulation", in Proceedings of the ACL '04, Barcelona, Spain, 2004.

[3] J.D. Moore, M.A. Foster, O. Lemon, and M. White, "Generating Tailored, Comparative Descriptions in Spoken Dialogue", in Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, AAAI Press, 2004.

[4] M.A. Walker, S. Whittaker, A. Stent, P. Maloor, J.D. Moore, M. Johnston and G. Vasireddy, "Generation and evaluation of user tailored responses in dialogue", Cognitive Science, 28, 811-840, 2004.

[5] V. Demberg and J.D. Moore, "Information Presentation in Spoken Dialogue Systems", in Proceedings of the 11th Conference of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL '06), Trento, Italy, 2006.

[6] J. Hu, A. Winterboer, C.I. Nass, J.D. Moore, R. Illowsky, "Context & usability testing: User-modeled information presentation in easy and difficult driving conditions", in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07), San Jose, CA, 2007.

[7] N. Dahlbaeck, A. Joensson and L. Ahrenberg, "Wizard of Oz Studies – Why and How", Knowledge-Based Systems, Vol. 6, No. 4, pp. 258-266, 1993.

[8] C.A. O'Muircheartaigh, J.A. Krosnick, A. Helic, "Middle alternatives, acquiescence, and the quality of questionnaire data", presented at the Ann. Meet. Am. Assoc. for Public Opin. Res., Fort Lauderdale, FL, 1999.