Evaluating the effect of information presentation strategies on task success and user perceptions

Andi Winterboer* and Johanna D. Moore*

Abstract. In this paper we present results of a Wizard-of-Oz (WoZ) experiment comparing two approaches to information presentation in spoken dialogue systems. We show that the user-model based summarize- and-refine (UMSR) approach leads to increased task efficiency in comparison with the summarize-and-refine (SR) approach. Moreover, user ratings on four evaluation criteria demonstrated a clear preference for recommendations based on the UMSR approach to information presentation.

1 INTRODUCTION

Spoken dialogue systems (SDS) are developed to provide natural and efficient access to information and carry out simple tasks using speech as the interaction mode. A common task for spoken dialogue systems is to help users select a suitable option (e.g., flight, hotel, restaurant) from the set of options available. When the number of options is small, they can simply be presented sequentially. However, as the number of options increases, the system must have strategies for helping users browse the space of available options. Thus, intelligent algorithms for effective information presentation must be devised. In this paper, we present a user study that compares two previously proposed approaches to information presentation in terms of their effect on task success and user perceptions.

2 RELATED WORK

Recently, a number of approaches to information presentation have been proposed. In presentations based on the *summarize and refine* (SR) approach, introduced by [1] and later extended by [2], when the number of options to be presented is large, the system structures them into a small number of clusters that share attribute values. The system then summarizes the clusters based on their attributes and prompts the user to provide additional constraints in order to narrow down the number of options. Thus, the system supports the user by reducing the large number of options to a small number of clusters. 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, [3] identified several limitations to this approach. First, depending on the size and complexity of the domain, many turns may be required during the refinement process. Because the user's preferences are not taken into account, the clusters may contain many irrelevant entities, which must be filtered out successively with each refinement step. Second, if there is no optimal solution, exploration of trade-offs between options can be difficult. Finally, the chosen clusters may be based on attributes that are irrelevant for the specific user. Figure 1 shows an example dialogue based on the SR approach.

User:	I would like to go from San Francisco to Prague on June 14th.
SR:	I found 21 flights from San Francisco to Prague. All of 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, please.
SR:	I found 11 business class flights from San Francisco to Prague. All these flights require a connection. There are flights in the moderate and expensive price range. I also have information about arrival times.
User:	I'd like to arrive before 3 p.m.
SR:	…

Figure 1 SR dialogue sample

In contrast, 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 [4,5]. Although the UM approach to information presentation enables a concise presentation of a small number of options, highlighting the ways in which those options satisfy user preferences, it does not scale up to presenting a large number of options. When there are hundreds of potentially relevant options to consider (e.g., all Italian restaurants in New York City), there may be a large number of options that fit the user model, making it difficult to narrow down the number of options. In addition, users may not be able to provide constraints until they are presented with more details concerning the option space. Moreover, the system does not provide an overview of the available options, which may lead to the user's actual or perceived missing out on possibly better alternatives.

The user-model based summarize and refine (UMSR) approach to information presentation was devised in an attempt to combine the benefits of the UM and SR approaches [3]. It employs a user model to reduce dialogue duration by considering only options that are relevant to the user. When the number of relevant items exceeds a manageable number, the UMSR approach builds a cluster-based tree structure which orders the options for stepwise refinement based on the ranking of attributes in the user model. The effectiveness of the tree structure, which directs the dialogue flow, is enhanced by taking the user's preferences into account. Trade-offs between alternative options are presented explicitly in order to provide the user with a better overview of the option space. In addition, to give users confidence that they are being presented with all of

* School of Informatics, Univ. of Edinburgh, Edinburgh, EH8 9AB, UK. Email: {A.Winterboer, J.Moore}@ed.ac.uk.

the relevant options, a brief account of all the remaining (irrelevant) options is also provided. Thus, the UMSR approach maintains the benefits of user tailoring, while allowing for presentation of large numbers of options in an order reflecting user preferences. Figure 2 shows an example dialogue based on the UMSR approach to information presentation.

User:	I'd like to go from San Francisco to Prague on June 14th, please.
UMSR:	There are no direct flights from San Francisco to Prague, but I found 11 flights with availability in business class. Two of these are on KLM.
User:	Are there any flights arriving before 3 p.m?
SR:	There are two business class flights 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:	Please book the second flight.

Figure 2 UMSR dialogue sample

Previous work has shown that users significantly prefer the UMSR approach [3]. In that study participants were asked to read transcripts of six dialogue pairs, each consisting of one SR dialogue and one UMSR dialogue. After reading each dialogue, participants were asked to judge it according on the following four criteria:

1. *Understandability* ("Did the system give the information in a way that was easy to understand?"),
2. *Overview of options* ("Did the system give the user a good overview of the available options?"),
3. *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
4. *Efficiency* ("How quickly did the system allow the user to find the optimal flight?")

In this study, users rated the UMSR approach significantly more highly on criteria 2-4, and found the UMSR and SR approaches equally easy to understand [3]. The study was replicated using an "overhearer" technique in which, rather than reading transcriptions of dialogues, participants listen to dialogues between a "user" and a simulated spoken dialogue system. Again, participants showed a significant preference for dialogues in which the UMSR approach to information presentation was used.

In subsequent work, we conducted two user studies to examine the impact of the two different information presentation methods on a secondary task, namely driving [6,7]. In these experiments, participants interacted with what they thought was a spoken dialogue system, and thus we were able to assess the impact of the two approaches on effectiveness criteria such as task duration and completion. We found that: (1) the UMSR approach enables more efficient information retrieval in comparison to SR approach (requiring fewer turns and shorter dialogue duration), the UMSR approach was more effective (i.e., users were significantly more likely to pick the flight that best matched the user model) than the SR approach, and that presenting information with UMSR did not negatively affect driving performance.

However, in contrast to results of the previous, showing significant preferences for UMSR when participants were reading or overhearing dialogues, no differences between user satisfaction ratings of the two presentation methods were observed in the dual task studies. Thus, in order to find out whether the lack of differences between the user satisfaction ratings was caused by the fact that participants were actually conversing with a SDS (as opposed to simply reading or "overhearing" the dialogues), or whether the reason was the demanding secondary task, we performed the following experiment in which participants exclusively interacted with the simulated SDS (instead of performing an additional task simultaneously).

3 USER STUDY

For the current experiment, the same information retrieval and presentation methods were used as in the dual task experiments [6,7]. A total of 34 participants, all naïve to the purpose of the experiment and mostly students of the University of Edinburgh, were paid to participate in the experiment. The average age of the 17 female and 17 male participants was 24.26 years. The Wizard-of-Oz (WoZ) methodology [8] provides the opportunity to test hypotheses about not yet implemented systems, such as complex spoken dialogue systems, by simulating the system.

For the experiment, participants sat facing a wall, in front of a desk equipped with a laptop computer, two microphones, and small speakers. The wizard sat on the opposite side of the room, hidden behind a partition that prevented participants from seeing or hearing the wizard during the experiment. The wizard's computer was connected to the speakers and the microphones on the participant's desk via cables running on the floor along the walls of the room in order to not attract attention.

A database-driven Web interface was used to generate system responses on-the-fly based on either the SR or the UMSR strategy to information presentation. The wizard performed speech recognition and natural language understanding. The integrated SQL-based database contained actual flight information as provided by airlines (taken from www.expedia.co.uk). The wizard used drop-down menus to perform stepwise queries according to participants' requests until a satisfactory flight was found and booked. The generated textual information provided by the Web interface was then synthesized by Speechify™, a text-to-speech application provided by Nuance Communications, Inc. Because prior research showed that people clearly identify with and prefer a voice that "matches" their gender [9], all participants heard a synthetic voice of their own gender. Participants were encouraged to speak naturally rather than merely responding to system prompts. Therefore, the wizard used very few questions as prompts and would only add additional questions if the participant remained silent for more than five seconds after each round of information presentation by the system.

To begin the experiment, participants were asked to read the instructions on the computer screen explaining that they would

be booking four flights with a spoken dialogue system. In order to enable reliable and rigorous comparisons, all participants were instructed to play the role of a business traveller for the flight booking task. In descending order of importance, the business traveller 1) prefers flying *business class*, 2) is (equally) concerned about *arrival time*, *travel time*, and *number of stops*, and 3) wants to *fly on KLM* if possible. 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 randomized to counter-balance possible order effects. The order of the deployed information presentation strategies was also alternated. Half of the participants obtained flight information presented from using the SR approach; the other half received search results presented with the UMSR approach.

The following experimental phase consisted of two major steps. In Step 1, the participant was informed that she would interact with a "flight information system" to book a total of four flights. She was requested to pretend she was a "business traveller" 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 traveller's motivation to travel to the specific destination.

In the second step, the wizard started the conversation with the first system utterance: "This is the flight information system. I am now connected to the network. Would you like to book a flight?" A conversation began as soon as the participant responded to this prompt. The wizard performed database queries and converted textual output into synthetic speech using the text-to-speech system. After confirming the second flight booking, participants filled in a questionnaire containing the evaluation questions introduced above. Additionally, the participant received instructions on booking two more flights. However, this time they received system utterances based on a different presentation method, i.e., participants who received SR-based presentations for the first two flights received UMSR-based presentations for the next two flights and vice versa. After completing the last of the four flights, the participant again received a user questionnaire to provide feedback on the four criteria introduced before. Then, the participant was debriefed, paid, thanked, and discharged.

4 RESULTS

Dialogues were recorded and analysed. Data captured by the questionnaires were tabulated and analysed in SPSS. For the questionnaire data, the same seven-point Likert scales as in [3] were used.

Overall, there was a highly significant difference in the number of turns each participant required for booking a flight when the system adopted the SR approach in comparison with the UMSR approach to information presentation (see Figure 1). Participants using UMSR took significantly fewer turns (10.53) than when using the SR based search system (14.53, $p < .0001$).

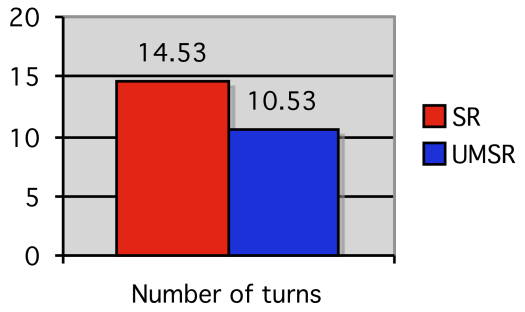

Figure 1: Number of turns per booking with SR and UMSR

Furthermore, there was a highly significant difference in the average dialogue duration between bookings made with presentations based on UMSR and SR. When the system presented information using the UMSR approach, participants were able to complete their task in significantly less time (see Figure 2, SR=391.65s, UMSR=252.55s, again $p < .0001$).

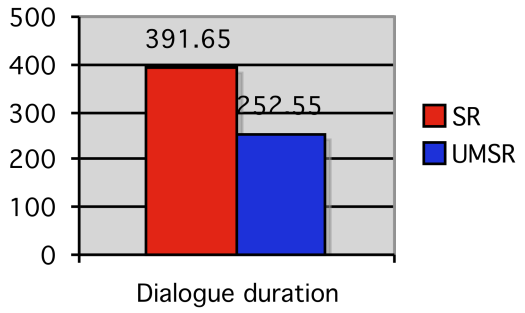

Figure 2: Average dialogue duration for two bookings with SR and UMSR (in seconds)

In addition, we had hypothesized that UMSR, which explicitly points out trade-offs among options, would lead to improved task success. In order to test this hypothesis, we counted how often the flight "best" matching the business traveller's profile was chosen in each condition. As can be seen in Figure 3, we found again a significant difference between the two conditions ($p < .05$). Sixty-eight "best" flights could be booked with each system. However, with presentations based on the SR approach only 50 "best" flights were booked, in comparison to 62 with presentations based on UMSR.

In summary, the average flight booking process with a system using recommendations based on UMSR had considerably shorter dialogue duration and required fewer dialogue turns. Moreover, the flight best matching the business traveller's profile was selected significantly more often in the UMSR condition in comparison with the SR condition. Therefore, in terms of task success and dialogue efficiency the findings of [7] were replicated.

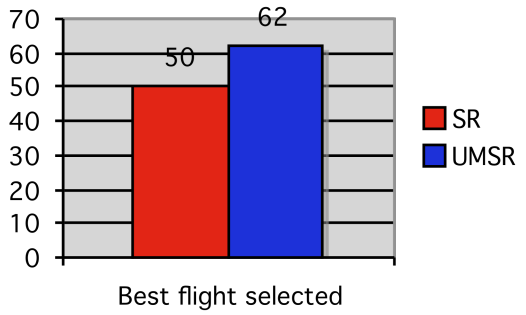

Figure 3: How often was the best flight selected?

In addition, in the obtained questionnaire data we found a general preference for UMSR-based recommendations on all four evaluation criteria (introduced in Section 2). However, only differences between answers to the first ("Did the system give the information in a way that was easy to understand?", $p < .05$), and last question ("How quickly did the system allow the user to find the optimal flight?", $p < .005$) were statistically significant (see Table 1).

	Understandability	Overview of options	Relevance of options	Efficiency
SR	5.27*	4.85	3.76	4.86*
UMSR	5.79*	5.18	4.00	5.63*

Table 1: User satisfaction ratings (7 Likert-type scale)

5 DISCUSSION

The results of prior experiments, which asked participants to evaluate presentations based on SR and UMSR presented as dialogue transcripts [3] or as sound files demonstrated a clear preference for presentations based on UMSR. We did not find the same preferences when another highly demanding secondary task was conducted simultaneously in a dual task experiment [6,7], in which participants interacted with a simulated dialogue system. However, the 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 or ability to answer them accurately. The current user study was carried out to determine whether this lack of difference was due to the extensive evaluation questionnaire or the cognitive load imposed by the demanding driving task, which potentially influenced the participants' perception of the system. The results of this experiment seem to suggest that the secondary task did affect user ratings. In this experiment, where participants interacted with simulated systems implementing the two different information presentation strategies, we found that presenting information with UMSR not only leads to greater task success, considerably shorter task duration, and requires fewer turns, but users also preferred UMSR over SR presentations. Possibly, participants in conditions of high cognitive load are so concerned with completing the dual tasks that they are less aware of differences in the order in which options and their attributes are presented, or the wording of the presentation. However, it is important to keep in mind that in both dual task experiments [6,7] UMSR was more effective in terms of task success and dialogue efficiency (i.e., task duration, number of turns).

6 CONCLUSIONS AND FUTURE WORK

In this paper we presented the results of a Wizard-of-Oz experiment comparing two approaches to information presentation in spoken dialogue systems. In line with results from previous experiments [6,7] we found that in terms of task success and dialogue duration the user-model based summarize and refine (UMSR) approach clearly outperforms the summarize and refine (SR) approach, and enables more effective information retrieval and information presentation. In addition, we also found user ratings on the four user satisfaction criteria to demonstrate a consistent trend favouring presentations based on the UMSR approach.

We are also interested in understanding just what it is about the presentations generated by the UMSR approach that makes it superior to the SR approach. We hypothesize the benefits arise because the use of a user model allows the generation of presentations that explicitly point out trade-offs among the options, rather than requiring the user to compute the trade-offs mentally. To test this hypothesis, we recently performed an additional experiment in which participants read messages that presented information about consumer products (hotel rooms, restaurants, microwaves, Mp3 player). We compared two different information presentation message types. In the first type, the trade-offs among options are made explicit with linguistic devices (e.g., discourse cues, adverbials). These presentations are similar to the presentations produced by the UMSR approach to information presentation reported in this paper. In the second type, trade-offs are not made explicit, and thus they are similar to the presentations based on the SR approach. In this experiment, we were interested in item recall. We found that item recall was significantly greater when the texts included linguistic cues highlighting properties of and relations between items (e.g., trade-offs) [9]. We are currently running a web-based experiment, in which participants hear the two different types of presentations and are again asked recall questions, in order to determine if the finding in [9] holds for spoken presentations.

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, 2003.

[2] G. Chung. Developing a flexible dialogue system using simulation. In: Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics *(ACL '04)*, Barcelona, Spain, 2004.

[3] V. Demberg and J.D. Moore. Information presentation in spoken dialogue systems. In: Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics *(EACL '06)*, Trento, Itlay, 2006.

[4] J.D. Moore, M.E. Foster, O. Lemon, and M.White. Generating tailored, comparative descriptions in spoken dialogue. In: *Proceedings of FLAIRS '04*. AAAI Press, 2004.
[5] 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.
[6] J. Hu, A. Winterboer, C.I. Nass, J.D. Moore, and R. Illowsky. Context & usability testing: User-modelled information presentation in easy and difficult driving conditions. In: Proceedings of Computer/Human Interaction Conference *(CHI '07)*, San Jose, CA, 2007.
[7] A. Winterboer, J. Hu, J.D. Moore, and C.I. Nass. The influence of user tailoring and cognitive load on user performance in spoken dialogue systems. In: Proceedings of Interspeech '07, Antwerp, Belgium, 2007.
[8] N. Dahlbaeck, A. Joensson, and L. Ahrenberg. Wizard of Oz studies – Why and How. *Knowledge-based systems*, 6(4):258-266. (1993)
[9] A. Winterboer, J. D. Moore, and F. Ferreira. Do discourse cues facilitate recall in information presentation messages? In: Proceedings of Interspeech '08, Brisbane, Australia, 2008.