Evaluating Information Presentation Strategies for Spoken Recommendations

Andi Winterboer
University of Edinburgh
School of Informatics
Edinburgh, United Kingdom
A.Winterboer@ed.ac.uk

Johanna D. Moore
University of Edinburgh
School of Informatics
Edinburgh, United Kingdom
J.Moore@ed.ac.uk

ABSTRACT

We report the results of a Wizard-of-Oz (WoZ) study comparing two approaches to presenting information in a spoken dialogue system generating flight recommendations. We found that recommendations presented using the user-model based summarize and refine (UMSR) approach enable more efficient information retrieval than the data-driven summarize and refine (SR) approach. In addition, user ratings on four evaluation criteria showed a clear preference for recommendations based on the UMSR approach.

Categories and Subject Descriptors: H.5.2 [Information Interfaces and Presentation]: User Interfaces – natural language, evaluation/methodology, user-centered design; H.1.2 [Models and Principles]: User/Machine Systems – human factors, human information processing; I.2.7 [Artificial Intelligence]: Natural Language Processing – language generation; H.3.3 [Information storage and retrieval]: Information Search and Retrieval.

General Terms: Algorithms, Experimentation, Human Factors.

Keywords: Information presentation, spoken dialogue systems, user modeling.

1. INTRODUCTION

As information becomes abundant, and access to it becomes more important, we face the problem of choosing between all the available options. *Recommendation systems* [1] are intended to guide users through the (potentially large) space of available options and support users in selecting the most suitable option. Most recommendation systems, in research as well as in commercial use, are text- and graphics-based, and presentation of options is done using a display, where users can see many options at a time (when using desk or laptop computers) and/or scroll back (as when viewing options on a mobile phone) to view previously mentioned options. However, there is also great interest in information services that are intended for situations where users' eyes and hands are busy, e.g., in-car voice services, in which user preferences must be elicited and recommendations made using spoken natural language as the main interaction mode. Indeed, spoken dialogue systems (SDS) have been developed to provide information about a wide range of products and services, e.g., restaurant recommendation [7, 10], and travel information [6]. The naturalness and perceived intelligence of a spoken dialogue interface does not depend only on its ability to recognise and analyse user utterances correctly, but also on the quality of the information presentation module. Recent work on evaluating SDS has shown that dialogue duration is negatively correlated with user satisfaction [9], and our own analysis of the DARPA Communicator Corpus has shown that the information presentation phase is the primary contributor to dialogue length. Therefore, it is crucial that we gain understanding of how best to design information presentation strategies for spoken recommendations. In this study, two previously introduced approaches to information presentation are compared in terms of their impact on effectiveness and user satisfaction.

In recommendations based on the summarize and refine (SR) approach, developed by [7] and later extended by [2], a large number of options is grouped 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, [4] argued that there are limitations to this approach. First, many turns may be required during the refinement process depending on the size and complexity of the domain. Second, if there is no optimal solution, exploration of trade-offs is difficult. Finally, the chosen clusters may be irrelevant for the specific user. This approach to information presentation is comparable to that of typical *conversational recommender systems* [8] which utilize conversations between the user and the system to determine and constantly update user preferences during the dialogue to arrive at a recommendation best matching the user's preferences. However, in contrast to such systems, the SR approach does not create a user model. Table 1 shows an example dialogue based on the SR approach.

The user-model based (UM) approach [6, 10] employs a user model and decision theory techniques to identify and present a small number of options that best match the user's preferences. Even though this approach to information presentation enables a concise presentation of a small number of options, highlighting the ways in which those options sat-

RecSys'07, October 19–20, 2007, Minneapolis, Minnesota, USA.
Copyright 2007 ACM 978-1-59593-730-8/07/0010 ...$5.00.

Table 1: *SR dialogue sample*

User:	"I would like to go from San Francisco to Prague on June 14^{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 would 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 expensive and and moderate price range. I also have information about arrival times."
User:	"I'd like to arrive before 3 p.m."
SR:	"..."

Table 2: *UMSR dialogue sample*

User:	"I'd like to go from San Francisco to Prague on June 14^{th}, please."
UMSR:	"There are no direct flights from S. Francisco to Prague, but I found 11 flights with availability in business class. 2 of these are on KLM."
User:	"Do you have 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."

isfy user preferences, it does not scale up to presenting a large number of options. When there are hundreds of potentially relevant options to consider, there may be many that fit the user model. Additionally, users may not be able to provide constraints until they are presented with more details concerning the option space. Finally, since the UM based approach does not provide an overview of the available options, it may lead to the user's actual or perceived missing out on possible better alternatives.

More recently, [4] proposed the user-model based summarize and refine (UMSR) approach, which combines the benefits of the UM and SR approaches. The UMSR approach to information presentation 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. In order to provide the user with a better overview of the option space, trade-offs between alternative options are presented explicitly. We hypothesize that this also allows the user to make a more informed choice. Finally, to give users confidence that they are being presented with all relevant options, a brief account of the remaining (irrelevant) options is also provided. Thus, the UMSR approach maintains the benefits of user tailoring, while allowing for presentations of large numbers of options in an order reflecting user preferences. Table 2 presents an example dialogue employing the UMSR approach.

In an earlier study, researchers found significant preferences for the UMSR approach when participants read transcripts of dialogues [4]. To obtain the user judgments, 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. Participants were asked to judge each dialogue on the following 4 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?").

[4] found that users rated the UMSR approach significantly more highly on criteria 2-4, and found the UMSR and SR approaches equally easy to understand. This 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 SDS [Moore, p.c.]. Again, users preferred the UMSR approach.

In subsequent work, we carried out 2 studies to examine the impact of the two different information presentation methods on a secondary task, namely driving [5, 11]. In these experiments, participants actually interacted with what they thought was a spoken dialogue system, and thus we were able to assess the impact of the different approaches on effectiveness criteria such as task duration and completion. We found that the UMSR approach enables more efficient information retrieval in comparison to the summarize and refine approach, and that presenting information with UMSR did not negatively affect driving performance. However, in contrast to results of the previous studies [[4], Moore, p.c.] 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 "overhearing" or reading the dialogues), or whether the reason was the demanding secondary task, we conducted the following experiment in which participants only interacted with the simulated SDS.

2. USER STUDY

For the current experiment, the same information retrieval and presentation algorithms were deployed as in the dual task experiment [11]. A total of 34 participants, 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. All participants were naive to the purpose of the experiment.

2.1 Experimental setup

The experiment was conducted in rooms of the University of Edinburgh. The participants sat 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 laptop computer was connected to the speakers and the microphones on the participant's desk via long cables running on the floor along the walls of the room in order to not attract attention.

2.2 Wizard environment

The Wizard-of-Oz approach [3] provides the opportunity to test hypotheses about not yet implemented systems, such as complex spoken dialogue systems, by simulating the system. For this study, a database-driven Web interface was used which generated system responses on-the-fly based on either the SR or the UMSR strategy to presenting information. The wizard performed speech recognition and natural language understanding. Furthermore, the wizard kept the dialogue going if the user was silent. The integrated SQL-based 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. The generated textual information provided by the Web interface was copied-and-pasted 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 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.

2.3 Procedure

Each participant was directly led to a chair in front of a table facing a wall. Then, they were asked to read the instructions on the laptop computer's 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 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 concerning the two flights to be booked in the first part of the experiment. 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 strategy was rotated as well. 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 in the second part of the experiments.

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 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 wizard started the conversation with the first system utterance: "This is the flight 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. The wizard performed database queries and converted textual output into synthetic speech. After confirming the booking of the second flight, the participant received a questionnaire containing the above introduced evaluation questions. 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 receiving 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 questionnaire to provide judgments on four criteria introduced above. Then, the participant was debriefed, paid, thanked, and discharged.

3. RESULTS

Dialogues were recorded and analyzed. Data captured by the questionnaires were tabulated and analyzed in SPSS. For the questionnaire data, the same seven-point Likert scales as in [4] were used.

3.1 Dialogue efficiency

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 to the system adopting the UMSR approach to information presentation (see Table 3). Participants using UMSR took significantly fewer turns than when using the SR-based search system ($p < .0001$, indicated with "**").

Table 3: *Number of turns per booking with SR and UMSR*

	SR (N=34)	UMSR (N=34)
Turns	14.53**	10.53**

In addition, there was a highly significant difference in the average dialogue duration between bookings made with presentations based on UMSR and SR. When the system used presentations based on the UMSR approach, participants were able to complete their task in less time (see Table 4).

We had hypothesized that UMSR, which explicitly points out trade-offs among options, would lead to improved task success. To test this hypothesis, we counted how often the flight "best" matching the business traveler's profile was chosen in each condition. Table 5 shows that there is a significant difference ($p < .05$) between the two conditions. Sixty-eight "best" flights could be booked with each system.

Table 4: *Average dialogue duration for 2 bookings with SR and UMSR*

	SR (N=34)	UMSR (N=34)
Duration (sec)	391.65**	252.55**

However, with presentations based on the SR approach only 50 "best" flights were booked in comparison to 62 with presentations based on UMSR.

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

	SR (N=34)	UMSR (N=34)
Best flight selected	50 (73.53%)*	62 (91.18%)*

Overall, the average flight booking process with a system using recommendations based on UMSR had considerably shorter dialogue duration and required fewer dialogue turns. In addition, the best available flight was selected significantly more often in the UMSR condition. Thus, information access with the UMSR approach is more efficient than with the SR approach. Therefore, in terms of task success and dialogue efficiency the findings of [11] were replicated.

3.2 User satisfaction ratings

In the obtained questionnaire data, presented in Table 6, we found a general preference for UMSR-based recommendations on all four evaluation criteria (introduced in Section 1). 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.

Table 6: *Answers to the 4 user satisfaction/evaluation questions (on a scale from 1-7)*

	understand-ability	overview of options	relevance of options	efficiency
SR	5.27*	4.85	3.76	4.86*
UMSR	5.79*	5.18	4	5.63*

4. DISCUSSION

The results of prior experiments, which asked participants to evaluate presentations based on SR and UMSR presented as dialogue transcripts [4] or as sound files [Moore, p.c.] demonstrated a clear preference for the UMSR strategy. We did not find the same preference when another highly demanding task was conducted simultaneously in a dual task experiment [11], in which participants interacted with a simulated dialogue system. The current user study was carried out to determine whether this lack of difference was due to the cognitive load imposed by the driving task which influenced the participants' perception of the system. The results of this experiment seem to suggest that the secondary task did affect user ratings. Possibly, participants in conditions of high cognitive load are so concerned with completing the dual tasks that they are less aware of differences in wording of presentations or differences in the order in which options and their attributes are presented. However, it is important to keep in mind that in the dual task experiments, UMSR was more effective in terms of task success and dialogue duration.

5. CONCLUSIONS

This paper presents results of a WoZ experiment comparing two different approaches to information presentation in spoken dialogue systems. In line with results from previous experiments [5, 11] 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. In addition, we also found user ratings on the four user satisfaction criteria to demonstrate a consistent trend favoring recommendations based on the UMSR approach.

6. REFERENCES

[1] R. Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[2] G. Chung. Developing a flexible spoken dialog system using simulation. In *Proc. of ACL*, Barcelona, Spain, 2004.

[3] N. Dahlbaeck, A. Joensson, and L. Ahrenberg. Wizard of oz studies – why and how. *Knowledge-Based Systems*, 6(4):258–266, 1993.

[4] V. Demberg and J. D. Moore. Information presentation in spoken dialogue systems. In *Proc. of EACL*, Trento, Italy, 2006.

[5] J. Hu, A. Winterboer, C. I. Nass, J. D. Moore, and R. Illowsky. Context & usability testing: User-modeled information presentation in easy and difficult driving conditions. In *Proc. of CHI*, San Jose, CA, 2007.

[6] J. D. Moore, M. E. Foster, O. Lemon, and M. White. Generating tailored, comparative descriptions in spoken dialogue. In *Proc. of FLAIRS*. AAAI Press, 2004.

[7] J. Polifroni, G. Chung, and S. Seneff. Towards the automatic generation of mixed-initiative dialogue systems from web content. In *Proc. of Eurospeech '03*, pages 193–196, Geneva, Switzerland, 2003.

[8] C. Thompson, M. Goker, and P. Langley. A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research (JAIR)*, 21:393–428, 2004.

[9] M. A. Walker, R. J. Passonneau, and J. E. Boland. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Proc. of ACL*, pages 515–522, 2001.

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

[11] 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 *Proc. of Interspeech 2007*, Antwerp, Belgium, 2007.