Context & Usability Testing: User-Modeled Information Presentation in Easy and Difficult Driving Conditions

Jiang Hu[1], Andi Winterboer[2], Clifford I. Nass[1], Johanna D. Moore[2], and Rebecca Illowsky[1]

[1]Department of Communication
Stanford University
Bldg. 120, 450 Serra Mall, Stanford, CA 94305
{huj, nass, rebeccai}@stanford.edu

[2]School of Informatics
University of Edinburgh
2 Buccleuch Place, Edinburgh, UK
{a.winterboer, j.moore}@ed.ac.uk

ABSTRACT

A 2x2 enhanced Wizard-of-Oz experiment (N = 32) was conducted to compare two different approaches to presenting information to drivers in easy and difficult driving conditions. Data of driving safety, evaluation of the spoken dialogue system, and perception of self were analyzed. Results show that the user-modeled summarize-and-refine (UMSR) approach led to more efficient information retrieval than did the summarize-and-refine (SR) approach. However, depending on driving condition, higher efficiency did not always translate into pleasant subjective experience. Implications for usability testing and interface design were presented, followed by discussions of future research directions.

Author Keywords

Information presentation, spoken dialogue system, user modeling, driving simulator, context of use, usability testing.

ACM Classification Keywords

H.5 Information interfaces and presentation; H.5.2 User-centered design; I.2.1 Natural language interfaces.

INTRODUCTION

A common task for spoken dialogue systems (SDS) 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. In this paper, we compare two recently proposed approaches in terms of the cognitive resources they require.

In the conventional summarize and refine (SR) approach [1, 4], the system groups a large number of options into a small number of clusters that share attributes. The system summarizes the clusters based on their attributes and then prompts the user to provide additional constraints. In the SR approach, 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. The drawbacks of this approach, however, include the large number of dialogue turns required for the refinement process and the possibility of irrelevant/uninformative clustering.

In the user-model (UM) based approach, the system identifies and presents a small number of options that best match the user's preferences [3, 5]. Although the UM approach may work well with a relatively small number of alternatives, it does not scale up to presenting tens or hundreds of options. In addition, the system does not provide an overview of options, which may lead to the user's actual or perceived missing out on potentially better options.

Recently, an alternative approach that combines the benefits of the two previously introduced approaches was proposed and studied [2]. In this user-modeled summarize and refine (UMSR) approach, the system exploits information from a user model to improve dialogue efficiency by 1) selecting options that are relevant to the user, and 2) introducing a content structuring algorithm that supports stepwise refinement based on the ranking of attributes in the user model. In this way, UMSR aims to keep the benefits of user tailoring, while extending the approach to handle presentation of large numbers of options in an order that reflects user preferences. Dialogue samples based on the SR and UMSR approaches can be found under the Experiment section.

In a previous laboratory experiment that compared the relative effectiveness of UMSR and SR approaches, participants read and evaluated transcripts of six manually-generated dialogue pairs based on both SR and UMSR [2]. Four criteria were used for the evaluation: understandability, overview of options, relevance of options, and efficiency. While the results clearly favor the UMSR approach, it is unclear whether such preferences would still be observed when the user is actually interacting with an SDS, and when the user is simultaneously conducting another task, such as driving a car.

CHI 2007, April 28–May 3, 2007, San Jose, California, USA.
Copyright 2007 ACM 978-1-59593-593-9/07/0004...$5.00.

In-Car Application of SDS

An important venue to deploy SDS is in vehicles because using eyes and hands for a secondary task may hinder driving [e.g., 5]. Not only do busy people have a growing need for information services while driving, but increasingly automakers also regard providing such services as a potential profit source. However, concerns over safety dictate the careful development and deployment of in-car applications of SDS. Presenting information to drivers requires the consideration of the distractive factor imposed by communicating with the SDS. This is especially true when driving conditions are unfavorable and demand a large portion of cognitive resources. In the meantime, cognitive load associated with driving may negatively impact the efficiency of interaction with the SDS.

Based on the rationale behind the UMSR approach, one would expect that, compared to an SR-based SDS, a UMSR-based SDS should a) be more efficient, b) cause fewer harmful distractions to drivers, and c) lead to more pleasant user experience, especially under difficult driving conditions. To test these hypotheses, the following lab experiment was designed and conducted.

EXPERIMENT

The experiment had a 2x2 mixed design. The style of information presentation (SR vs. UMSR) was a within-participant factor; cognitive load (easy vs. difficult driving course) was a between-participants factor.

Simulation of Driving

We use the STISIM Drive^{TM} simulation system and projected visuals on a wall-sized back-projection screen. A total of four courses with two levels of difficulty were used to vary driving-related cognitive load imposed on participants. With speed limits ranging from 25 mph to 55 mph, each course contained four sequential sections: a residential area, a small city, a country highway, and a big city. 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. Pre-tests proved that the difficult courses were harder to drive than the easy courses in terms of effects on actual and perceived driving performance. No significant difference was found between the two easy courses or the two difficult courses.

The simulator kept track of each participant's driving performance in terms of numbers of collisions, speeding tickets, traffic light and stop sign violations, and minor driving errors including centerline crossing and road edge excursion.

Simulation of In-Car Information System

An enhanced Wizard-of-Oz method was used to simulate the SDS. The wizard used a database-driven Web interface to generate natural language responses with either the SR or the UMSR approach. The algorithms were based on the persona described below and were similar to those described in [2]. The database contained actual flight information as provided by airlines. When the system adopted the SR approach to presenting information, the wizard used drop-down menus to perform stepwise queries upon request from participants until they found the satisfactory flight and made the booking decision. With the UMSR approach, the Web interface returned search results based on a business traveler's persona, but allows for additional stepwise refinement as well.

The wizard copied and pasted textual output from the Web interface 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 make requests for refinement rather than merely responding to system prompts. Consequently, 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.

Participants

A total of 32 students from Stanford University, all licensed drivers, were paid to participate in the study. Participants with prior exposure to driving simulator were excluded; gender was balanced across conditions.

Persona and Flight Booking

To make reliable and rigorous comparisons, participants were asked to assume a business traveler's persona 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. The following offers a side-by-side comparison of first-round presentations for this persona:

> SR: *"I found 23 flights from New York to Frankfurt. There are direct flights as well as flights that require a connection. I also have information about fare classes."*
>
> UMSR: *"I found 6 direct business class flights from New York to Frankfurt. None are on KLM. However, if you're willing to make a connection, there is a business class KLM flight arriving at 1:35 p.m., connecting in Amsterdam."*

Each participant drove for two experimental rounds and booked four different one-way flights. Prior to each round of driving, participants received detailed instructions on the two flights to be booked. To make the booking process more realistic, the four routes (i.e., pairs of cities) were carefully chosen so 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.

Procedure

Participants were randomly assigned to the "easy-driving" or the "difficult-driving" condition. The order of each participant's two courses was also randomized. During the first round of experimental driving, half of the participants received flight information presented with the SR approach; the other half heard search results presented with the UMSR approach. The opposite approach was used during the second round of experimental driving. The order in which the four flights were booked was rotated to counter-balance possible order effects.

Participants first drove on the demo course to familiarize themselves with the simulator. The experimental phase that followed consisted of three major steps. In Step 1, the participant was told to use an "in-car information system" to book flights while driving. She was instructed to assume the persona of the business traveler for the booking tasks. At the same time, she received instructions on booking the first two flights.

In Step 2, the participant drove on the first experimental course alone in the lab. About three minutes later, a short beep was played, followed by the first utterance from the system 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 sent by the wizard sitting in a neighboring room. Via a wireless connection, 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 wirelessly to speakers put near the simulator. After booking the first flight, the participant was prompted to book the second flight.

In Step 3, the experimenter returned to the lab and administered a questionnaire that asked the participant to evaluate the "in-car information system," herself during the interaction, and the driving condition. Ten-point Likert scales were used except for the four seven-point Likert scales from the previous study [2]. The ten-point scales were meant to capture subtle variations and to avoid a middle point that often encourages "satisficing."

Once the participant indicated that she was ready for the second round of driving, Steps 1 through 3 were repeated, with different flights to book, and a different course (of the same degree of difficulty) to drive. Upon completing the last questionnaire, the participant was debriefed, paid, thanked, and discharged.

RESULTS

Dialogues were recorded and transcribed; data captured by the driving simulator and the questionnaires were tabulated. Factor analyses were performed for all questionnaire items to extract reliable and meaningful indices. All indices are reliable with Cronbach's alpha values ranging from .65 to .92. A series of SPSS repeated-measure ANOVAs were conducted, followed by *post hoc* analyses when necessary.

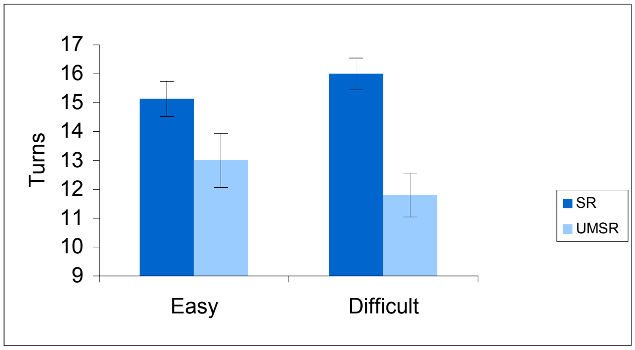

Figure 1. Average number of dialogue turns taken by participants to book two flights.

Manipulation Check

The manipulation of driving condition was successful. Specifically, although the average *number of collision accidents* was quite low, difficult-driving participants had significantly more accidents than easy-driving participants, $F(1,30) = 26.69$, $p < .001$, $M_{easy} = 0$ (0), $M_{difficult} = .82$ (.95). This was also true for the average *number of minor driving errors*, including center-line crossing and road edge excursions, $F(1,30) = 18.73$, $p < .001$, $M_{easy} = .60$ (.60), $M_{difficult} = 2.19$ (1.65). No difference was found in terms of stop sign and traffic light violations, and number of speeding tickets. Moreover, easy-driving participants rated their courses as much *easier* than did difficult-driving participants, $F(1,30) = 18.93$, $p < .001$, $M_{easy} = 7.63$ (.89), $M_{difficult} = 5.92$ (1.70).

Dialogue Efficiency

Participants in general took fewer dialogue turns when the system adopted the UMSR approach than when it utilized SR, $F(1,30) = 19.96$, $p < .001$, as shown in Figure 1; the average duration of dialogue (in seconds) was also shorter when the system used the UMSR than the SR approach, $F(1,30) = 8.58$, $p < .01$, $M_{SR} = 465.85$ (84.20), $M_{UMSR} = 401.00$ (108.48). These results supported Hypothesis a).

Driving Safety

Participants had significantly more minor errors when the system adopted the UMSR approach than when it used SR, $F(1,30) = 6.08$, $p < .05$, $M_{SR} = 1.09$ (1.45), $M_{UMSR} = 1.69$ (1.49), but this appears to be driven by the difference observed among easy-driving participants. Hypothesis b) was not supported. In fact, the reverse was true for easy-driving participants. However, their average number of minor errors was less than one, thus having little negative impact on driving safety.

Perceptions

System: There appeared to be a cross-over interaction between driving condition and the style of information presentation on the participants' perception of how *fun* the system was, $F(1,30) = 7.24$, $p < .05$. *Post hoc* analyses suggest that easy-driving participants thought that the UMSR approach was more fun to use than was the SR, and difficult-driving participants were more likely than easy-driving participants to think that the SR was fun to use.

Answers to the four questions/scales used in the previous study [2] were also analyzed. The only significant result was that participants thought that UMSR was more likely than SR to overlook better options, $F(1,30) = 5.33$, $p < .05$, $M_{SR} = 3.94$ (1.93), $M_{UMSR} = 4.68$ (1.67), but this difference was primarily observed among difficult-driving participants.

Self: Overall, the participants perceived themselves more *positively*[1] when the system adopted the SR approach to present search results, $F(1,30) = 9.65$, $p < .01$. Once again, this main effect appeared to be driven by the difference observed among difficult-driving participants.

An interaction of the presentation style and driving condition was found on participants' self-reported *friendliness*,[2] $F(1,30) = 7.44$, $p < .05$. *Post hoc* analyses indicate that easy-driving participants thought that they were friendlier when the system adopted the UMSR approach than when it adopted the SR approach, and they were more likely than difficult-driving participants to perceive themselves as friendly when the system presented information with the UMSR approach.

The above subjective findings were mixed; Hypothesis c) was partially supported.

Finally, a comparison of participants' self-reported usual driving behavior and in-experiment driving behavior shows an interaction between driving condition and presentation style, $F(1,30) = 6.25$, $p < .05$. Specifically, easy-driving participants reported that they had reduced offensive driving (suggesting more cautious driving) when the system had adopted the SR approach, and had increased offensive driving when it presented information in the UMSR style. There was also an expected main effect of driving condition, such that difficult-driving participants drove more cautiously than did easy-driving participants.

DISCUSSION

Although there was a slight increase in minor driving errors when the system used the UMSR approach as opposed to the SR approach, the general finding is that voice-based browsing using UMSR is more efficient than one that adopts the SR approach. This is consistent with the findings of [2], and provides behavioral evidence supporting the UMSR approach.

However, improved dialogue efficiency with an SDS does not necessarily lead to positive subjective user experience. In our study, only participants in easy driving conditions were able to appreciate UMSR's tailored presentations, despite the high efficiency of UMSR in all conditions. Whereas participants in the previous study believed that UMSR provides better overview than does SR [2], our participants thought otherwise when driving conditions were unfavorable. Findings like this unequivocally highlight the importance of context of use in usability testing, and prompt researchers to identify problems with interface design.

A further examination of transcribed dialogue files helped us uncover a critical flaw with our current UMSR simulation: for one of the four city pairs, the system generated an extremely long first-round presentation with a user-modeled summary followed by details of three flights. Moreover, there were redundant and unnecessary pieces of information within that long presentation. Even though the presentation was tailored for the persona, the large amount of information nonetheless placed a huge cognitive burden on our participants, especially when driving-related cognitive load was already heavy.

In theory, UMSR systems should be intrinsically superior to SR systems because they leverage knowledge of the user. The key challenge, then, is to utilize the strength of UMSR systems without burdening the user with too much information. If this goal can be achieved, using an in-car SDS can be made safer, more efficient, *and* more pleasant.

REFERENCES

1. Chung, G. Developing a flexible spoken dialog system using simulation. In *Proc. of ACL 2004*, ACL (2004), 63-70.
2. Demberg, V., and Moore, J.D. Information presentation in spoken dialogue systems. In *Proc. of EACL 2006*, ACL (2006), 65-72.
3. Moore, J.D., Foster, M.E., Lemon, O., and White, M. Generating tailored, comparative descriptions in spoken dialogue. In *Proc. of the 17th International Florida Artificial Intelligence Research Society Conference*, AAAI Press (2004), 917-922.
4. Polifroni, J., Chung, G., and Seneff, S. Towards automatic generation of mixed-initiative dialogue systems from web content. In *Proc. of Eurospeech '03*, 193-196.
5. Salvucci, D.D. Predicting the effects of in-car interface use on driver performance: an integrated model approach. *Int. J. Human-Computer Studies*, *55*, 85-107.
6. Walker, M.A., Whittaker, S., Stent, A., Maloor, P., Moore, J.D., Johnston, M., and Vasireddy, G. Generation and evaluation of user tailored responses in dialogue. *Cognitive Science*, *28*, 811-840.

[1] This index is composed of 10 items such as *competent*, *powerful*, *skilled*, *successful*, and *intelligent*.

[2] This index is composed of three scales: *cooperative*, *friendly*, and *polite*.