

USER-TAILORED GENERATION FOR SPOKEN DIALOGUE: AN EXPERIMENT

Amanda Stent, Marilyn Walker, Steve Whittaker, Preetam Maloor

AT&T Labs - Research
180 Park Ave
Florham Park, NJ, USA, 07932
stent,walker,whittaker,pmaloor@research.att.com

ABSTRACT

Recent work on evaluation of spoken dialogue systems suggests that the information presentation phase of complex dialogues is often the primary contributor to dialogue duration. Therefore, better algorithms are needed for the presentation of complex information in speech. This paper evaluates the effect of a user model on generation for three dialogue strategies: SUMMARY, COMPARE and RECOMMEND. We present results showing that (a) both COMPARE and RECOMMEND strategies are effective; and (b) the user model is useful.

1. INTRODUCTION

Recent work on evaluating spoken dialogue systems suggests that the information presentation phase of complex dialogues is often the primary contributor to dialogue duration [10]. During this phase the system returns from a database query with a set of options that match the user's constraints. The user then navigates through these options and refines them by offering new constraints. Depending on the number of options returned, and the facilities provided for refinement, this process may be quite onerous.

In related work [9], we describe algorithms that support this refinement process for the task of selecting restaurants in MATCH (Multimodal Access to City Help) [5]. Unlike most previous work on natural language generation, these algorithms are designed for the presentation of speech, as opposed to text, output. To address user problems of remembering complex spoken data, our algorithms are based on an underlying *user model* enabling us to present information that is directly relevant to specific user preferences. Furthermore, the SUMMARY, RECOMMEND and COMPARE dialogue strategies we developed allow us to characterize the general properties of returned option sets, and to highlight the options and attributes that are most relevant to choosing between them, again addressing user's problems in remembering complex spoken information.

This paper presents an evaluation of these strategies for information quality for both speech and text presentations. In particular, we test whether providing an underlying user model increases the perceived usefulness of generated dialogue responses. We ask subjects to evaluate the information quality of responses tailored to them, versus responses tailored to a randomly selected other subject. Our hypothesis is that if the user models are different enough, there should be differences in the subjects' perceptions of the responses. We also test, for each response, the perceived quality of the response presented in speech and in text. While our algorithms are designed for speech presentation, a subject's perception of the spoken response can be confounded by the quality of the text-to-speech in the restaurant domain.

Section 2 describes how we construct and elicit user models. After explaining our general procedure and providing example user mod-

els, we describe how we quantify the notion of DISTANCE between user models that we use to test our hypothesis. Section 3 briefly describes the SPUR (Speech Planning with Utilities for Restaurants) generator and the MATCH system in which SPUR is a module. Section 4 describes our experimental design and Section 5 presents our results. Our results show that (1) the user models need only be slightly different to have an effect on the user's perception of information quality; (2) the textual presentation of responses is preferred over the spoken presentation. User comments suggest that this is because the names of restaurants are hard to understand. We wait until Section 6 to discuss related work and provide conclusions.

2. CONSTRUCTING AND ELICITING USER MODELS

Multi-attribute decision models are based on the fundamental claim that if anything is valued it is valued for more than one reason [6]. In the restaurant domain, this implies that a user's preferred restaurants optimize a combination of restaurant attributes. In order to define a multi-attribute decision model for the restaurant domain, we must determine these attributes and their relative value for particular users.

Edwards and Barron describe a procedure called SMARTER for eliciting multi-attribute decision models for particular users or user groups [4]. First, the important attributes in the domain, and their relationships to each other, are identified. Second, the values of each attribute are mapped to single-dimension cardinal utilities that span the whole range from 0 to 100. Third, a function is defined that combines the utilities for each attribute into an overall utility score for an option. Finally, weights (or rankings) are assigned to each attribute that indicate the importance of that attribute to each user. The SMARTER procedure also specifies how to elicit these weights from users in a way that takes little time and has been shown to result in more accurate user models than simple ranking [4].

Attribute	Range of values	Mapping of values to cardinal utilities
Food quality, Service, Decor	0-30	value x 3 1/3
Cost	0-90	100 - (10/9 x value)
Food type, Neighborhood	e.g. Italian, French, West Village	top values listed by user are mapped to 90, bottom ones to 10 and all others to 50

Table 1. Mapping of attribute values to utilities in the restaurant domain

Table 1 shows the attributes in the restaurant domain, with the functions mapping the values of each attribute to cardinal utilities.

Each attribute is assumed to be independent of every other one. The individual attribute utilities are combined into an overall utility using a simple additive function; the value for each attribute is multiplied by its weight and all the weighted values are summed.

The user models for the subjects in our experiment were collected in a separate process from the experiment itself. User model elicitation was done over the web, and consisted of a series of very simple questions designed to elicit attribute rankings. The resulting user model is stored in a separate file for each subject. It contains numerical values for the relative importance (ranking) of all six attributes and categorical values for food type and location. While most users included cost and food quality in their top three attributes, the relative importance assigned to other attributes, such as location and food type, varied widely. Two sample user models are shown in Figure 2.

3. INCORPORATING USER-ADAPTED GENERATION INTO MATCH

We implemented our information presentation strategies in the context of AT&T's MATCH multimodal dialogue system. MATCH permits users to obtain information about restaurants in New York city and find subway routes between locations. It consists a collection of heterogeneous components, including speech recognition, parsing, dialogue management, interface management and text-to-speech components, that communicate using message-passing through a central hub. To enable MATCH to produce user-tailored RECOMMEND, SUMMARY and COMPARE strategies, we added the SPUR generator and user modeling component to the system.

SPUR takes as input the following: (a) the user model for a particular user; (b) the database of restaurant information, which contains values for each restaurant for each of the six attributes in our domain; and (c) the situational constraints arising from the user's query, which select a particular set of restaurant options from the database. It uses the user model ranking of attributes to assign cardinal utilities (weighted values) for each of the 6 attributes to each restaurant, and to rank the restaurants by overall utility.

SPUR also uses information from the user model to select appropriate content for each restaurant to be described. The actual process of content selection is different for each strategy, as described in more detail in [9]. SPUR produces a set of content items and the rhetorical relations holding between them [7]. At the moment, we do not have a sentence planner or surface generator in MATCH, so SPUR also does some simple template-based surface generation and produces a text based on the selected content.

Table 2 shows user models for two subjects. Table 3 shows the top-ranked restaurants for a selection of restaurants for each of these user models. Because CK values food quality most highly, while OR values cost most highly, the restaurant rankings differ for each subject.

Because of the necessity for brevity in speech output, when recommending a restaurant or comparing among several restaurants, the system only describes those restaurants that are most highly ranked, i.e. that most closely match the user's preferences. The user model also guides the choice of what to say about each option. For example, *Uguale* is highly ranked for OR because of its cost. It seems reasonable, therefore, to mention that attribute before mentioning service, which hardly contributes at all to the overall ranking of *Uguale*.

Table 4 illustrates the effects of the user model on these two tasks, i.e. ranking sets of options returned from a database query and choosing what to say about each selected option (different restaurant selections were made for each strategy). In this work, we are interested in exploring these two effects of the user model. We do this by compar-

ing subjects' judgments about dialogue strategies generated using (a) the subject's own user model and (b) a user model for someone else. By randomly selected another user model (Random) we can explore the question of how *different* two user models have to be to make a difference in the subject's perception of system responses. In order to quantify the notion of *different* we define a measure of DISTANCE between two user models. To compute DISTANCE we sum, over all attributes, the absolute values of the differences between the rankings for each attribute. For example, the distance between models CK and OR in Table 2 is 0.84. The average distance between two models for the subjects in our experiment is 0.56.

We also wanted to determine how well our strategies addressed the problems of speech output, and so we compared user judgments for speech and text output.

4. EXPERIMENTAL DESIGN

The experimental procedure we followed is similar to that used in the SPoT experiments [8]. The subject is an "overhearer" of a series of dialogues, each involving one restaurant-selection task. In each dialogue, output for all three strategies is presented on separate web pages. There are four tasks in this experiment, each involving one or two constraints: (a) French restaurants; (b) restaurants in Midtown West; (c) Italian restaurants in the West Village; (d) Asian restaurants in the Upper West Side. The tasks were chosen to accommodate a variety of user models, to provide sets of restaurants large enough to be interesting and to be fairly easy for subjects to remember.



Fig. 1. User circles subset of Italian West Village restaurants for comparison.

Each web page sets up the task by showing the MATCH system's graphical response for an initial user query, e.g. *Show Italian restaurants in the West Village*. Then the page shows the user circling some subset of the restaurants and asking the system to *summarize, compare or recommend* options from the circled subset. Figure 1 shows an example of the user circling a set in order to ask for a comparison.

The subject sees one page each for SUMMARY and RECOMMEND, and two for COMPARE, for each task. On each page, the subject sees one system response tailored to her user model, and a different one tailored to the user model of another randomly selected subject. The order of the tasks, and the order of appearance of strategies within the task, does not vary between subjects. However, the order of presentation of subject-tailored and other-tailored responses varies at random from page to page.

User	Food Quality	Service	Decor	Cost	Nbhd	FT	Nbhd Likes	Nbhd Dislikes	FT Likes	FT Dislikes
CK	0.41	0.10	0.03	0.16	0.06	0.24	Midtown, Chinatown, TriBeCa	Harlem, Bronx	Indian, Mexican, Chinese, Japanese, Seafood	Vegetarian, Vietnamese, Korean, Hungarian, German
OR	0.24	0.06	0.16	0.41	0.10	0.03	West Village, Chelsea, Chinatown, TriBeCa, East Village	Upper East Side, Upper West Side, Uptown, Bronx, Lower Manhattan	French, Japanese, Portugese, Thai, Middle Eastern	no-dislike

Table 2. Sample User Models

Name	Utility	Food-Q (WTD)	Service (WTD)	Decor (WTD)	Cost (WTD)	Neighborhood (WTD)	Food Type (WTD)
CK							
Babbo	66	26 (36)	24 (8)	23 (2)	60 (5)	W. Village (3)	Italian (12)
Il Mulino	66	27 (38)	23 (7)	20 (2)	65 (4)	W. Village (3)	Italian (12)
Uguale	64	23 (29)	22 (7)	18(2)	33 (11)	W. Village (3)	French, Italian (12)
Da Andrea	60	22 (26)	21 (6)	17 (1)	28 (12)	W. Village (3)	Italian (12)
John's Pizzeria	59	22 (26)	15 (3)	13 (1)	20 (14)	W. Village (3)	Italian, Pizza (12)
OR							
Uguale	69	23 (17)	22 (4)	18 (9)	33 (28)	W. Village (9)	French, Italian (2)
Da Andrea	69	22 (16)	21 (4)	17 (8)	28 (31)	W. Village (9)	Italian (1)
John's Pizzeria	68	22 (16)	15 (2)	13 (5)	20 (35)	W. Village (9)	Italian, Pizza (1)
Vittorio Cucina	64	22 (16)	18 (3)	19 (9)	38 (26)	W. Village (9)	Italian (1)
Babbo	62	26 (21)	24 (5)	23 (12)	60 (14)	W. Village (9)	Italian (1)

Table 3. Top five restaurants ranked by two user models for the query *Italian restaurants in the West Village*. Domain attribute values are given along with WTD = Weighted utility for that attribute.

For each instance of a RECOMMEND, SUMMARY, or COMPARE, the subject is asked to state her degree of agreement (on a scale from 1 to 5) with the following statement, intended to determine the informativeness, or *information quality*, of the response: *The system's utterance is easy to understand and it provides exactly the information I am interested in when choosing a restaurant.*

This entire sequence of web pages is presented twice. The first time through, the subject can only read (not hear) the system responses. The second time, she can only hear them. We used this read-then-hear approach in order to obtain subject ratings that are not biased by the performance of the text-to-speech.

To summarize, each subject "overhears" a sequence of four dialogues about different restaurant-selection tasks. The entire sequence is presented twice (once for text, once for speech). The subject makes eight information quality judgments for each dialogue each time. The total number of information quality judgments per subject is sixty-four. The total time required to complete the experiment is approximately half an hour per subject.

Sixteen subjects completed the experiment. All are fluent English speakers. Most eat out moderately often (seven eat out 3-5 times per month, six 6-10 times). All sixteen currently live in northern New Jersey. Eleven described themselves as somewhat or quite familiar with Manhattan, while five thought they were not very familiar with it. After the experiment, ten subjects identified themselves as very interested in using a system like MATCH in the future.

5. RESULTS

Our hypothesis is that if the user models are different enough, there should be differences between the subject's perception of the useful-

ness of responses tailored to her and those tailored to some other subject. We first tested whether any difference at all in the user model affected subjects' rankings of the information quality of the system's responses and whether this varied according to the strategy. A two-way ANOVA for information quality by strategy and model indicates no overall effect of model ($F = 1.4$, $p = 0.23$ n.s.). This result suggests that some level of DISTANCE between models is required before users can perceive differences between the system responses.

However, the ANOVA for information quality by strategy and model does indicate a significant difference between strategies ($F = 127.9$, $p = 0.0001$). An exploration of this difference shows that the SUMMARY strategy was clearly less effective than the other strategies (the mean score for summaries was 2.33; that for comparisons was 3.53; and that for recommendations was 4.08). User comments at the end of the experiment also qualitatively confirm this finding; many users commented that the summaries were too high-level to be useful.

We then turned to the question of how large the DISTANCE has to be between models to affect users' perceptions. We filtered the user judgments to compare only those responses generated in a context where the distance between user models was greater than 0.2. This eliminated 48 judgments leaving a set of 464 paired judgments. A paired t-test for information quality by user model over all strategies for this subset of the data indicates a significant effect of user model on information quality ($df = 463$, $t = 2.61$, $p = .009$). This result shows that user models only have to be *minimally* different to affect users' perceptions. As mentioned above, the average distance between user models in our sample is .57.

This experiment was also designed to test for differences between spoken and text presentation of user tailored responses. A one-way ANOVA of information quality by mode of presentation indicates significant differences ($df=1$, $F = 4.751$, $p = 0.030$). However, the subjects preferred the text presentations over the spoken ones. The mean

Speech plan type	Output
USER OR	
compare	<i>Among the selected restaurants, the following offer exceptional overall value. Ugualè's price is 33 dollars. It has good decor and very good service. It's a French, Italian restaurant. Da Andrea's price is 28 dollars. It has good decor and very good service. It's an Italian restaurant. John's Pizzeria's price is 20 dollars. It has mediocre decor and decent service. It's an Italian, Pizza restaurant.</i>
summary	<i>The 3 restaurants have the same price range of 28 to 33 dollars. They differ in food quality and decor.</i>
recommend	<i>Ugualè has the best overall value among the selected restaurants. Ugualè's price is 33 dollars.</i>
USER CK	
compare	<i>Among the selected restaurants, the following offer exceptional overall value. Babbo's price is 60 dollars. It has superb food quality, excellent service and excellent decor. Il Mulino's price is 65 dollars. It has superb food quality, excellent service and very good decor. Ugualè's price is 33 dollars. It has excellent food quality, very good service and good decor.</i>
summary	<i>The 3 restaurants have the same price range of 28 to 33 dollars. They differ in food quality.</i>
recommend	<i>Babbo has the best overall value among the selected restaurants. Babbo has superb food quality.</i>

Table 4. Sample output from the SPUR Generator

score of the text model responses was 3.45 and the mean score of the speech mode responses was 3.28. We hypothesize that this difference arose because the text-to-speech engine does not contain in-domain vocabulary for the restaurant domain; users commented that they often had trouble understanding the restaurant names.

6. CONCLUSIONS AND FUTURE WORK

Our work is a direct extension of research by Carenini and Moore, who built a generation system for constructing user-tailored text recommendations of houses [1, 2]. They showed that subjects are more satisfied with the house selections they make if they see tailored recommendations than if they see non-tailored recommendations or no recommendations at all, and that they also spend less time making selections if they see tailored recommendations [3].

We extended this work in two ways. First, we explored the effects of user modeling on the generation of more types of discourse; in particular, our generation system produces summaries and comparisons as well as recommendations. This requires organizing information about multiple options in one presentation. Second, we chose to focus on generation for spoken dialogue rather than generation of text. Speech is uniquely ephemeral. This makes it especially important that information presented in speech be carefully selected and organized.

Our results indicate that the usefulness of comparisons as well as recommendations can be improved if they are tailored to individual users. While we were not able to demonstrate the utility of user-tailored summaries, subject comments indicate that this is an artifact of the structure of our summaries, not a general principle. Our results also indicate the strong effect of text-to-speech performance on the perceived usefulness of system responses in spoken dialogues. In short, SPUR is helpful not only because it reduces the time required for users to examine a selection of restaurants; it also improves the chances that users will get useful, relevant information about restaurants they are likely to be interested in.

In the future we plan to conduct additional experiments in this framework. There are several variables in the content selection algorithm that were held constant for this experiment but that can significantly affect system responses. Among these are cut-off levels (such as discussing only the those attributes rated most important by the user) and outlier values (which determine how different a particular attribute's value must be from the norm for it to be considered "outstanding"). We also did not vary the overall text structures for each

strategy in this experiment, although in our own exploration we identified various possibilities for each strategy. Additional experiments will alter these constraints, and explore subject preferences for the resulting output.

7. REFERENCES

- [1] G. Carenini. *Generating and evaluating evaluative arguments*. PhD thesis, University of Pittsburgh, 2000.
- [2] G. Carenini and J. Moore. A strategy for generating evaluative arguments. In *Proceedings of the International Natural Language Generation Conference*, Mitzpe Ramon, Israel, 2000.
- [3] G. Carenini and J. Moore. An empirical study of the influence of user tailoring on evaluative argument effectiveness. In *Proceedings of IJCAI 2001*, Seattle, USA, 2001.
- [4] W. Edwards and F. Barron. SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement. *Organizational Behavior and Human Decision Processes*, 60:306–325, 1994.
- [5] Michael Johnston, Srinivas Bangalore, and Gunaranjan Vasireddy. MATCH: Multimodal access to city help. In *Automatic Speech Recognition and Understanding Workshop*, Madonna Di Campiglio, Trento, Italy, 2001.
- [6] R. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley and Sons, 1976.
- [7] Daniel Marcu. Building up rhetorical structure trees. In *Proceedings of AAAI/IAAI 1996*, volume 2, pages 1069–1074, 1996.
- [8] M. Walker, O. Rambow, and M. Rogati. SPoT: A trainable sentence planner. In *Proceedings of the North American Meeting of the Association for Computational Linguistics*, 2001.
- [9] Marilyn Walker, Steve Whittaker, Amanda Stent, Preetam Maloor, Johanna Moore, Michael Johnston, and Gunaranjan Vasireddy. Speech plans: Generating evaluative responses in spoken dialogue. In *Proceedings of the International Natural Language Generation Conference*, 2002.
- [10] Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems. In *Proceedings of the Meeting of the Association of Computational Linguistics*, 2001.