

Should I Tell All?: An Experiment On Conciseness in Spoken Dialogue

Stephen Whittaker, Marilyn Walker, Preetam Maloor

University of Sheffield, University of Sheffield, University of Toronto

s.whittaker@shef.ac.uk, walker@dcs.shef.ac.uk, pmaloor@cs.toronto.edu

Abstract

Spoken dialogue systems have a strong requirement to produce concise and informative utterances. While interacting over a phone, users must both understand the system's utterances, and remember important facts that the system is providing. Thus most dialogue systems implement some combination of different techniques for (1) *option selection*: pruning the set of options; (2) *information selection*: selecting a subset of information to present about each option; (3) *aggregation*: combining multiple items of information succinctly. We first describe how user models based on multi-attribute decision theory support domain-independent algorithms for both option selection and information selection. We then describe experiments to determine an optimal level of conciseness in information selection, i.e. how much information to include for an option. Our results show that (a) users are highly oriented to utterance conciseness; (b) the information selection algorithm is highly consistent with user's judgments of conciseness; and (c) the appropriate level of conciseness is both user and dialogue strategy dependent.

1. Introduction

Spoken dialogue systems have a strong requirement to produce concise and informative utterances. While interacting over a phone, users must be able to both understand the system's utterances, as well as remember important facts that the system is providing. Utterances that are too long, or that contain too much information will not be easily understood or remembered. This requirement is especially important during the information presentation phase of the dialogue. After obtaining enough user constraints, the system looks up potential options matching the constraints in a database of information, such as information about airline flights, train schedules, or entertainment options [1, 5]. Often, more options are returned than the system can provide in one system turn; it is also often the case that more information is available about each option than the system can provide in a single turn. Thus most dialogue systems implement some combination of different techniques for (1) *option selection*: pruning the set of options; (2) *information selection*: selecting a subset of information to present about each option; (3) *aggregation*: combining multiple items of information succinctly.

One technique for option selection is to base the pruning algorithm on a model of the user. In previous work, we describe how user models based on multi-attribute decision theory support an algorithm for the selection of options that are tailored to an individual user's known preferences [6, 7]. This selection procedure can be used to select a single best option and recommend that to the user, or to select a subset of options in order to provide the user with a comparison between several highly ranked options. We implemented these techniques in AT&T's MATCH system (Multimodal Access to City Help), a system that provides information about restaurant and entertainment options in New York City [3]. We showed experimentally that users preferred recommendations and comparisons that were tailored to their individual preferences. In these experiments, we also used the decision-theoretic user models for algorithms for information selection, i.e. given a particular level of conciseness, the user model determines which attributes of an option

Usr	Concise?	Output
CK	Concise	Bond Street has the best overall value among the selected restaurants. Bond Street has excellent food quality.
CK	Sufficient	Bond Street has the best overall value among the selected restaurants. Bond Street has excellent food quality. It's a Japanese, Sushi restaurant.
CK	Verbose	Bond Street has the best overall value among the selected restaurants. Bond Street's price is 51 dollars and it has excellent food quality and good service. It's a Japanese, Sushi restaurant.
BA	Concise	Komodo has the best overall value among the selected restaurants. Komodo's a Japanese, Latin American restaurant.
BA	Sufficient	Komodo has the best overall value among the selected restaurants. Komodo's price is 29 dollars and it has very good service. It's a Japanese, Latin American restaurant.
BA	Verbose	Komodo has the best overall value among the selected restaurants. Komodo's price is 29 dollars and it has very good service, very good food quality and good decor. It's a Japanese, Latin American restaurant.

Figure 1: Recommendations for Users CK and BA, for the East Village Japanese Task, of Varying Levels of Conciseness.

should be presented. These algorithms are domain-independent.

Figure 1 shows recommendations generated by MATCH for users BA and CK for the task of finding a Japanese restaurant in the East Village. The figure illustrates user-tailored option selection and information selection. Option selection is illustrated by the fact that different restaurants are recommended to each user. The known preferences for users BA and CK, as represented in their user models, leads the algorithm to select Bond Street for CK and Komodo for BA. Information selection is illustrated by the fact that, for the same level of conciseness, different attributes are mentioned to BA and CK, again according to their user models.

However, in previous work we did not attempt to determine (a) whether users were sensitive to the dimension of conciseness; and (b) an optimal level of conciseness. Figure 1 shows both verbose and concise descriptions for the same restaurant, for the same user, that mention many or few attributes. This paper presents and evaluates a technique for manipulating conciseness, based on multi-attribute decision theory. We experimentally evaluate users' sensitivity to conciseness and determine the level of conciseness users perceive as optimal for user-tailored recommendations and comparisons. We first motivate the algorithm, describe the evaluation method, and then present the experimental results.

2. An Algorithm for Information Selection

Multi-attribute decision models: Multi-attribute decision models assume that if anything is valued it is valued for multiple reasons [4]. In the restaurant domain, the approach assumes that a user's preferred restaurants optimize a combination of restaurant attributes. In order to define a multi-attribute decision model for this domain, we first determine these attributes and their relative value for particular users. This involves the identification of (a) important domain attributes; (b) derivation of the user model.¹

¹Our focus here is on the information selection component of our algorithm when applied to recommend and compare strategies. The

Usr	FQ	SVC	DEC	Cost	Nbhd	FT	Nbhd Likes	Nbhd Dislikes	FT Likes	FT Dislikes
CK	0.41	0.06	0.03	0.10	0.16	0.24	Upper East Side, Upper West Side, Chinatown, Lower Manhattan, TriBeCa SoHo	The Bronx, Harlem	Chinese, French, Indian, Mexican, Tex-Mex	Health Food, Noodle Shops, Thai, Vegetarian, Vietnamese
BA	0.10	0.16	0.06	0.24	0.03	0.41	Downtown, Midtown, East Village, TriBeCa SoHo	The Bronx, Harlem	Cajun Creole, Greek, Italian, Japanese, Seafood	Coffeehouses Desserts, German, Steak

Figure 2: Example User Models: FQ = Food Quality, SVC = Service, DEC = Decor, Nbhd = Neighborhood, FT = Food Type

User	Restaurant	U_h	FQ(wtd)	SVC(wtd)	DEC(wtd)	Cost(wtd)	Nbhd(wtd)	FT(wtd)
BA	Komodo	77	22(7)	22(10)	19(4)	29(18)	2	36
BA	Japonica	71	23(7)	18(7)	15(3)	37(16)	2	36
BA	Takahachi	71	21(6)	17(6)	14(2)	27(19)	2	36
BA	Shabu-Tatsu	70	20(5)	18(7)	15(3)	31(17)	2	36
BA	Bond Street	69	25(8)	19(8)	22(4)	51(11)	2	36
BA	Dojo	66	15(2)	12(2)	8(1)	14(23)	2	36
CK	Bond Street	63	25(34)	19(3)	22(2)	51(5)	7	12
CK	Japonica	59	23(29)	18(3)	15(1)	37(7)	7	12
CK	Komodo	59	22(26)	22(4)	19(2)	29(8)	7	12
CK	Takahachi	54	21(24)	17(2)	14(1)	27(8)	7	12
CK	Shabu-Tatsu	52	20(22)	18(3)	15(1)	31(7)	7	12
CK	Dojo	30	15(10)	12(1)	8(0)	14(10)	7	12

Figure 3: Results of DB Query for East Village Japanese for users BA and CK: U_h = Overall Utility. WTD = Weighted utility for each attribute. FQ = Food Quality, SVC = Service, DEC = Decor, Nbhd = Neighborhood, FT = Food Type

The important domain attributes correspond to attributes about restaurants in the domain database, which may have continuous or categorical values. The continuous attributes in the MATCH database are Food Quality, Service, Decor and Cost. The categorical attributes are Neighborhood and Food Type.

The user model is represented as a linear function; the overall utility of a restaurant option is the weighted linear sum of the attribute values, where the weights of each attribute are specific to each user. The real-domain values of each attribute x must first be transformed into single-dimension cardinal utilities $u(x)$ such that the highest attribute value is mapped to 100, the lowest attribute value to 0, and the others to values in the interval 0 to 100. Then, if h ($h = 1, 2, \dots, H$) is an index identifying the restaurant options being evaluated, k ($k = 1, 2, \dots, K$) is an index of the attributes, and w_k is the weight assigned to each attribute, the overall utility U_h is defined as follows.

$$U_h = \sum_{k=1}^K w_k u_k(x_{h,k})$$

Figure 2 shows example user models for users CK and BA. The column values are the weights assigned to each attribute in the user model or the category instance values for category preferences. Preferred values for categorical attributes are mapped to the high end of the utility scale and nonpreferred values to the low end. Note that, when selecting a restaurant, user CK places greatest emphasis on Food Quality, Food Type and Neighborhood, while user BA values Food Type, Cost and Service.

Figure 3 shows how the user model affects the results of a database query for “Japanese restaurant in the East Village”, for users BA and CK. The same set of restaurant options are returned, but their ranking by overall utility is user specific, as reflected in their ranking according to the U_h column in Figure 3. The attribute columns show both the attribute values, and the weighted attribute values after mapping into the utility scale and weighting by each user model (wtd).

The Conciseness Algorithm: The weighted attribute model enables us in principle to determine the likelihood that mentioning a given attribute will change the user’s belief state. For example, compare the recommendations in Figure 1. The most concise recommendation for both CK and BA mentions one attribute. The weighted attribute values for each user in

Figure 3 predict how convincing a recommendation would be that includes that attribute. Figure 3 indicates that telling CK about Bond Street’s Food Quality should provide 34 utils (units of utility) out of a possible 63. Similarly, telling BA about Komodo’s food type is predicted to provide 36 utils out of a possible 77. Including more attributes makes the recommendation more convincing, e.g. adding the Food Type attribute as in CK’s Sufficient recommendation in Figure 1 should provide 46 (34 + 12) utils out of a possible 63 total utils.

In sum, we map conciseness directly onto the weighted attribute ranking of the user model: more concise descriptions select the subset of attributes that maximally affect the user’s belief state. More verbose descriptions also include lower weighted attributes. Obviously, however, there is a trade-off between maximizing expected utility, and verboseness. Mentioning more attributes increases expected utility while requiring the user to remember more information. The algorithm and experiment described below explore these trade-offs: concise descriptions mention only the highest weighted attribute, sufficient descriptions mention the top three weighted attributes, and verbose descriptions mention 5 attributes.²

3. Experimental Method

The experimental procedure is adapted from previous experiments [6]. The subject is an “overhearer” of a series of dialogues, each involving one restaurant selection task. In each dialogue, the output for each strategy is presented on a separate web page. There are 6 tasks in the experiment, each involving one or two constraints: (a) cheap restaurants; (b) restaurants in Midtown West; (c) Italian restaurants in the West Village; (d) restaurants in the Upper West Side; (e) French restaurants; (f) Japanese restaurants in the East Village. The tasks were chosen after extensive piloting to accommodate a variety of user models, to be fairly easy for subjects to remember, and to provide sets of potential restaurants large enough to be interesting.

The user models for the subjects in our experiment were collected in a separate process that took place before 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. This is done as part of registering as a user of the dialogue system.

reader should refer to [6] for full details of the derivation of the underlying user models and how these are used in the strategies.

²If an attribute is mentioned in the query, it is filtered out unless the attribute value is only a partial match to the query, as in the BA examples in Figure 1.

Each web page set up the task by showing the MATCH system’s graphical response for an initial user query, e.g. “Show Japanese restaurants in the East Village”. Then the page showed the user circling some subset of the restaurants and asking the system to compare or recommend options from the circled subset. See Figure 4.

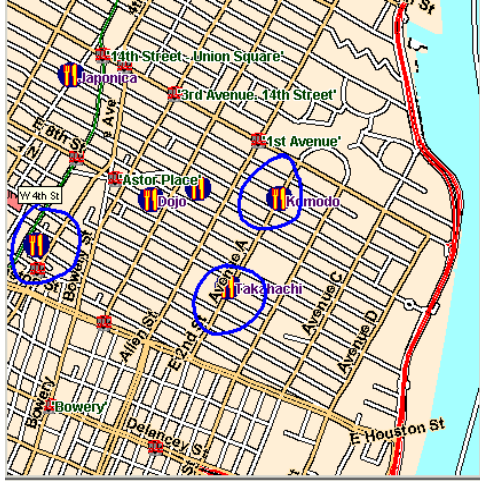


Figure 4: User interacting with map display in MATCH and circling a subset of Japanese Restaurants in the East Village.

Subjects saw one page each for recommend, and compare, for each task. On each page, they saw multiple system responses of differing conciseness. The order of the tasks, and the order of appearance of strategies within the task was consistent across subjects. However, the order of presentation of conciseness variants was randomized from page to page. For each instance of a recommend, or compare, the subject was asked to state her degree of agreement (on a 5-point Likert scale) with the following statement, intended to determine the conciseness of the response: “When choosing a restaurant, the amount of information provided by the system utterance is;” (1) far too little, (2) too little, (3) neither too little nor too much, (4) too much, (5) far too much.

To summarize, each subject “overheard” a sequence of 12 dialogues about 6 different restaurant-selection tasks (once for recommend, once for compare). The subject made 6 information quality judgments per task. The total number of information quality judgments per subject was 36. The total time required to complete the experiment was approximately half an hour per subject.

Twenty one subjects completed the experiment. All were fluent English speakers. We collected demographic information, about the frequency they ate out, and their familiarity with Manhattan, in case these affect their judgments. Most eat out moderately often (eleven eat out 3-5 times per month, ten 6-10 times). All subjects currently live in northern New Jersey or in Manhattan. Fourteen described themselves as somewhat or quite familiar with Manhattan, while seven thought they were not very familiar with it. After the experiment, 16 subjects (76%) identified themselves as agreeing with the statement that they would like to use a system like MATCH in the future.

4. Experimental Hypotheses

We tested two major hypotheses. Our first hypothesis addressed user’s sensitivity to conciseness and the correspondence between algorithmic conciseness and user judgments of conciseness. Our expectation was users would discriminate between different descriptions in terms of conciseness. More specifically, we expected that outputs we had operationalized as *concise* should be judged as providing too little information, out-

puts operationalized as *sufficient* should be judged as providing the right amount of information, and outputs operationalized as *verbose* should be judged as providing too much information.

Usr	Concise?	Output
CK	Concise	Among the selected restaurants, the top 3 in terms of overall value are as follows. Bond Street has excellent food quality. Japonica has excellent food quality. Komodo has very good food quality.
CK	Sufficient	Among the selected restaurants, the top 3 in terms of overall value are as follows. Bond Street has excellent food quality. It’s a Japanese, Sushi restaurant. Japonica has excellent food quality. It’s a Japanese, Sushi restaurant. Komodo’s price is 29 dollars and it has very good food quality. It’s a Japanese, Latin American restaurant.
CK	Verbose	Among the selected restaurants, the top 3 in terms of overall value are as follows. Bond Street’s price is 51 dollars and it has excellent food quality and good service. It’s a Japanese, Sushi restaurant. Japonica’s price is 37 dollars and it has excellent food quality and good service. It’s a Japanese, Sushi restaurant. Komodo’s price is 29 dollars and it has very good food quality and very good service. It’s a Japanese, Latin American restaurant.

Figure 5: Comparisons for User CK, for the East Village Japanese Task, of Varying Levels of Conciseness.

A second hypothesis concerned the relation between conciseness and information provision strategy. Contrast the comparisons in Figure 5 with those above in Figure 1. Of the two strategies, comparisons inherently contain more information than recommendations, because they mention multiple options and their attributes. We should therefore expect users to judge comparisons as more verbose.

We analyzed the user data using ANOVA. Independent measures were Algorithmic Verbosity (Verbose, Sufficient, Concise), Strategy (Recommend, Compare) and Task (cheap, Midtown West, West Village Italian, Upper West Side, French, East Village Japanese). We first transformed the elicited user judgments on a linear scale, so that an output judged to provide “far too little information” was scored -2, “too little” -1, “neither too much nor too little” 0, “too much” +1, and “far too much” +2. The transformed measure of conciseness was used as the ANOVA dependent measure.

Figure 6 indicates the relationship between Algorithmic conciseness and user judgments. It shows both that users are sensitive to conciseness and that user judgments paralleled our algorithmic implementation. Consistent with our hypothesis, outputs generated as concise were more likely to be judged as having too little information than those generated to be sufficient, which in turn were likely to have less information than those generated to be verbose ($F(2,1506)=559.1, p < 0.0001$), with post hoc tests showing judged differences between Algorithmically Concise and Sufficient, and between Algorithmically Sufficient and Verbose (both $p < 0.0001$). Furthermore, there was a strong correlation between algorithmic conciseness and the judgements ($r = .66, p < 0.0001$). Together these data show that we have algorithmic control over conciseness. Furthermore the algorithm was stable across tasks: there were no task differences or interactions involving task.

Nevertheless Figure 6 also indicates the need for further calibration of the algorithm. It shows that outputs generated to be sufficient are judged at -0.23, and those generated to be verbose are judged as 0.18. These observations suggest that we may be providing marginally too little information for our sufficient outputs and too little for our verbose outputs. This in turn would imply a need to tune the algorithm, in particular by adding more information to the sufficient statements. These data also suggest that judgments of conciseness may be asymmetric, indicating a preference for too much than too little information: even algorithmically verbose descriptions were judged as providing marginally too much information. This asymmetry may occur because users main concern is with whether they have enough information to carry out their task. Providing them with too little information is therefore judged negatively because they have

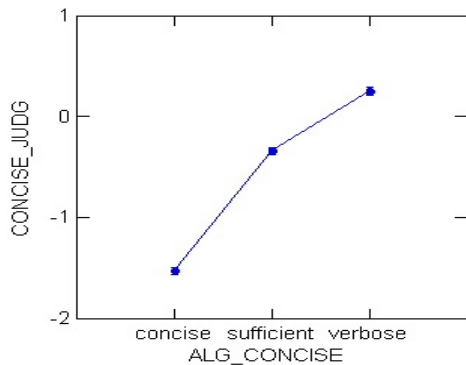


Figure 6: Relationship between Algorithmic Conciseness and User Evaluations.

insufficient information for the task. In contrast, providing too much information still allows them to execute the task, even though they have to remember more information to do this.

Our second hypothesis concerned the relationship between judged conciseness and strategy. Figure 7 shows as predicted that recommendations are judged to be more concise than comparisons ($F(1,1506)=4.4$, $p<0.05$). Furthermore, there is an interaction between strategy and judgments ($F(1,1506)=3.0$, $p<0.05$), with the main difference being accounted for by users' tendency to judge verbose comparisons as containing more information than verbose recommendations (post hoc test, $p < 0.05$). Possibly this was because verbose comparisons mention a total of 10 attributes, 5 more attributes than verbose recommendations, and this is perceived to be a large additional memory burden.

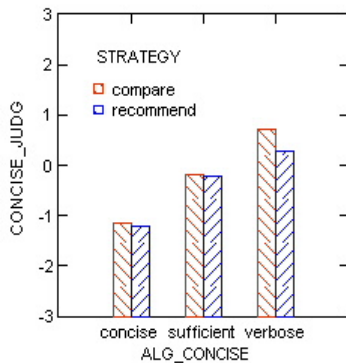


Figure 7: Conciseness Judgments for Different Strategies.

Finally, despite our tailoring of information content to individual users' preferences, it was clear that there were differences between users in terms of their overall perception of conciseness. Figure 8 shows users' judgments of overall presentation conciseness for each level of algorithmic conciseness. There are large individual differences between users. While most users judged that presentations provided slightly too little information overall, there was large individual variability, with some users judging presentations provided too little information (overall mean=-1.2), and others judging that presentation provided sufficient information (overall mean = 0.2). This suggests that the level of conciseness should also be tailored to individual preferences.

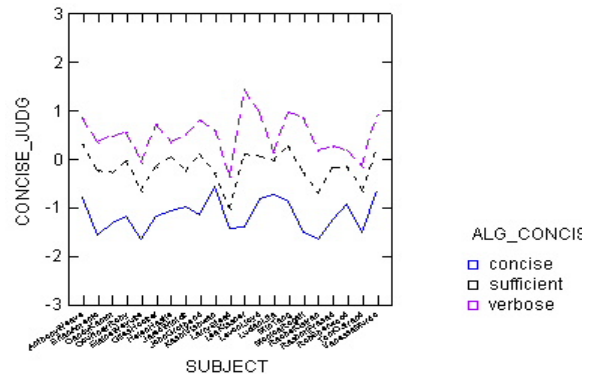


Figure 8: Individual differences in Conciseness Judgments.

5. Conclusions

We investigated techniques of information presentation attempting to optimize trade-offs between: (a) complete descriptions of complex information; (b) memory demands for remembering such information. We demonstrated the effectiveness of an algorithm intended to manipulate conciseness, although there are still outstanding issues concerning the exact calibration of this technique for different output strategies and users.

We build on related work by Carenini and Moore on generating user-tailored arguments in the real-estate domain [2]. They experimentally evaluated whether users preferred concise arguments over verbose arguments. Conciseness was manipulated by varying how many standard deviations a weighted attribute had to be away from the mean to be considered worth mentioning, i.e. how much of an outlier the weighted attribute value was. We also used this method in our previous work [6]. In the experiments reported here, we manipulated conciseness using a simpler method showing that users' judgements of conciseness were highly correlated with our algorithm manipulation.

6. Acknowledgements

Thanks to Michael Johnston and the other members of the MATCH team at AT&T Labs Research.

7. References

- [1] S. Bennacef, L. Devillers, S. Rosset, and L. Lamel. Dialogue in the railtel telephone based system. In *Proc. of ISSD*, pages 173–176, 1996.
- [2] G. Carenini and J. D. Moore. An empirical study of the influence of argument conciseness on argument effectiveness. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-00)*, 2000.
- [3] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. MATCH: An architecture for multimodal dialogue systems. In *Annual Meeting of the Association for Computational Linguistics, ACL*, 2002.
- [4] R. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley and Sons, 1976.
- [5] J. Polifroni and G. Chung. Promoting portability in dialogue management. In *Proc. of the 7th International Conference on Spoken Language Processing*, pages 2721–2724, 2002.
- [6] M. A. Walker, S. J. Whittaker, A. Stent, P. Maloor, J. D. Moore, M. Johnston, and G. Vasireddy. Speech-Plans: Generating evaluative responses in spoken dialogue. In *In Proc. of INLG-02.*, 2002.
- [7] S. Whittaker, M. Walker, and J. Moore. Fish or fowl: A wizard of oz evaluation of dialogue strategies in the restaurant domain. In *Language Resources and Evaluation Conference*, 2002.