Incorporating a User Model to Improve Detection of Unhelpful Robot Answers

Maxim Makatchev, Reid Simmons

***Abstract*— Dialogues with robots frequently exhibit social dialogue acts such as greeting, thanks, and goodbye. This opens the opportunity of using these dialogue acts for dialogue management, in particular for detecting misunderstandings. Our corpus analysis shows that the social dialogue acts have different scopes of their associations with the discourse features within the dialogue: greeting in the user's first turn is associated with such distant, or global, features as the likelihood of having questions answered, persistence, and ending with bye. The user's thanks turn, on the other hand, is strongly associated with the helpfulness of the preceding robot's answer. We therefore interpret the greeting as a component of a user model that can provide information about the user's traits and be associated with discourse features at various stages of the dialogue. We conduct a detailed analysis of the user's thanking behavior and demonstrate that user's thanks can be used in the detection of unhelpful robot's answers. Incorporating the greeting information further improves the detection. We discuss possible applications of this work for human-robot dialogue management.**

I. INTRODUCTION

Adapting to the model of a user's knowledge and/or emotional state (in short, a user model) has been shown to improve performance of dialogue systems on such metrics as time to task completion (e.g. [16]), learning gains [4], and learning efficiency and user's perception of the quality of the dialogue [3]. Deciding on the best action to be taken given a particular user and dialogue state can be done offline by analyzing a corpus of dialogues, but for online interaction, recognizing the state of the current user has to be performed on-the-fly from the ongoing interaction. For example the Roboceptionist [5], shown in Figure 1, is installed at a high-traffic entrance of a university building, and does not track users from session to session. Therefore, the user model has to be constructed from the first turns of the dialogue, so that the dialogue manager can take advantage of the model in adapting to the user while the interactive session is still in progress.

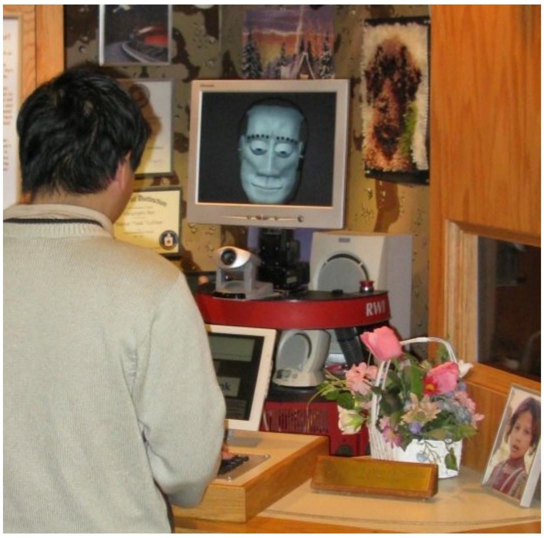


Fig. 1. Roboceptionist interacting with a user.

Our particular setting is additionally complicated by the fact that user input is typed, not spoken, thus excluding the possibilities of using prosody and other features of the user's speech. We analyzed transcripts of human-robot dialogues with the goal of finding dependencies between the first turn of the dialogue and dialogue patterns that potentially indicate a user's interaction style. In particular, we are interested in predicting such user traits as the willingness to carry on after the robot demonstrates lack of understanding (*Persistence*), as well as the user's adherence to social obligations, such as thanking the robot after the robot gives an answer to the user's question (*AnswerThanked*) and ending an interaction with a goodbye (*EndingWithBye*). We present results of the analysis of an annotated corpus of dialogues that demonstrates two significant associations. First, *EndingWithBye* and *QuestionAnswered* patterns are significantly associated with whether the dialogue was initiated by the human or by the robot. Second, the patterns of *Persistence*, *EndingWithBye*, *QuestionAnswered* and *AnswerThanked* are significantly associated with whether the user greeted the robot in the user's first turn. In particular, the considerable spans (in number of turns) observed between the *Greeting* and the patterns such as *AnswerThanked* and *EndingWithBye* motivate us to consider whether these associations are components of a user model that persists through the interaction.

As an application of the found associations, we focus on the problem of recognizing a specific type of the dialogue error, namely, the situation when the robot's answer is unhelpful to the user. We deem the robot's answer unhelpful if either the robot admits the failure to make sense of the question or to find the right piece of information, or the answer provided is irrelevant due to a misunderstanding at the intention and conversation levels [13], as in the example shown in Figure 2. Recognizing the unhelpfulness due mis-

This publication was made possible by the support of an NPRP grant from the Qatar National Research Fund. The statements made herein are solely the responsibility of the authors. The authors would like to thank Antonio Roque, Minh Hoai Nguyen, and Brett Browning for their comments.

Maxim Makatchev and Reid Simmons are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA {mmakatch,reids}@cs.cmu.edu

understanding is harder, because the degree of relevance of the answer is ultimately up to the user's interpretation. Previous work in detecting misunderstanding primarily targets automated speech recognition errors (e.g. [8], [10]) and has shown considerable accuracy using local dialogue context and prosodic features. Detecting unhelpfulness of an answer due to misunderstanding, in situations similar to the one shown in Figure 2, appears to require incorporating cues from the following user turns. For example, user corrections [15] and lexical-level features of the following user turn [6], [7] demonstrate good performance in detecting misunderstandings due to speech recognition errors.

The human-robot dialogues that we analyze regularly exhibit social dialogue acts, such as *Greetings*, *Thanks*, and *Goodbye*. We hypothesize that the social dialogue acts can be used as positive and negative cues and demonstrate that the user's *Thanks* is a significant predictor of the helpfulness of the robot's preceding answer. Unhelpful answer detection can be further improved by leveraging the difference in the thanking behavior between users that greeted the robot and users that did not.

Misunderstood questions can also be thought of in terms of grounding [2], defined as the process of adding material to the common ground between speakers. A misunderstood question in these terms is a type of error in adding material to the common ground between the user and the robot. Recognizing and recovering from a misunderstanding can therefore be viewed as a type of common ground maintenance. Managing the dialogue to achieve and maintain the desired degree of groundedness has been demonstrated to improve human perceptions of dialogues [14]. The system described in [14] predicts the degree of groundedness of a material by treating relevant dialogue acts as evidence of understanding, e.g. acts that acknowledge understanding, acts that refer to the material, or acts that rely on the understanding of the material. Our work is similar to the approach of [14] in that we use the following user's turn (namely the presence of *Thanks*) as a feature in our predictor of the answer helpfulness (which implies question understanding). Unlike [14], however, we also utilize the distant discourse information of the presence of *Greeting* in the user's first turn. In summary, we are modeling how well the answer was grounded, how well the question was understood, and how well the answer will be grounded (based on the presence of a *Greeting*). In this respect, the work on detecting unhelpful answers can be viewed as modeling actual and anticipated grounding behavior in human-robot dialogues.

The paper is organized as follows. Section II introduces the corpus of human-robot dialogues. Section III presents the results of the analysis of associations between the initiator of the interaction and discourse and between the presence of *Greeting* in the user's first turn and discourse. We analyze user's thanking behavior in Section IV. In Section V we evaluate the performance of the user's *Greeting* and user's *Thanks* as predictors of the answer unhelpfulness. The paper concludes with a discussion of the results and an outline of the future work on improving the user model and the ways it may be used in dialogue management.

II. HUMAN-ROBOT DIALOGUE

A. *Roboceptionist*

The Roboceptionist is a robot stationed in a kiosk at a high-traffic entrance of a Carnegie Mellon University building (Figure 1). The robot's face is rendered on a flat-screen display that is mounted on a neck joint enabling it to pan to follow the passers-by who are detected by the laser range scanner. Greeting of a passer-by is triggered by a user entering an area that is close to the robot with a minimal forward velocity. Regardless of whether the Roboceptionist has initiated the interaction by greeting the user, the user can start interacting with the robot by typing on the keyboard mounted in front of the robot. The Roboceptionist will respond by producing a synthesized voice reply as well as text that appears on the screen, next to its face. The first version of the Roboceptionist was introduced in 2003, and after an initial peak, it currently averages 30-40 interactions per day.

The robot has been provided with a back story covering its past career and personal life, developed by students in the Drama department. Occasionally, the robot refers to these story lines, which makes them a recurring topic of dialogues. The other topics that the robot is designed to handle, and that make a large fraction of the dialogues, include the weather and directions to rooms, buildings and people [9]. With the exception of followup questions, like the one in the second line of the dialogue in Figure 2, the robot's dialogue manager is reactive, namely it keeps track of only the last user turn. In particular, the robot does not remember whether it exchanged greetings with the current user.

B. *Human-robot dialogue corpus*

The corpus of human-robot dialogues that we analyze represents transcripts of uncontrolled interactions that are collected on a near-daily basis. To generate the dataset, the transcripts are first automatically segmented into individual dialogues, and dialogues with more than 20 turns[1] are discarded to eliminate outliers and some of the errors of the segmentation procedure. We annotated the dialogue using a multi-dimensional annotation in the spirit of DIT++ [1], tailoring it to our domain as necessary. Specifically, we manually labeled 1960 turns of 287 dialogues that occurred over 8 days in March of 2008 with respect to such dialogue acts as *Greeting*, *Thanks*, *Goodbye*, *UserQuestion*, *Answer*, *InterpNegFeedback* (robot's admitting its failure to make sense of the preceding user's turn), and *Rude* language. Discourse patterns of interest to us, such as *QuestionAnswered*, *Persistence*, *AnswerThanked* are expressed in terms of these dialogue acts.

We used this manually labeled corpus of dialogues to train decision tree classifiers for each of the dialogue acts. Using unstemmed words as the features, these classifiers

[1] Here, and in the remainder of the paper, we count separately each user and robot turn.

```
U: where can I find frank
R: Frank Jones is in 4001 Wean Hall. Would you like directions?
U: no
R: OK
U: I was lookinf for another frank
R: I didn't understand that. It might help if you rephrase it. [...]
U: I was looking for another frank
R: The closest match I have in my directory is Frank Jones in 4001 Wean Hall.
   Would you like directions?
U: any frank in smith hall
R: I didn't understand that. It might help if you rephrase it. [...]
U: camonn1
R: Whatever
```

Fig. 2. A verbatim fragment of an actual dialogue, with the person's last name and room modified for privacy, parts of robot utterances truncated for brevity, and the labels "U:" and "R:" added to denote the user and robot turns.

each achieve the accuracy of at least 89% (10-fold cross-validation is used to select the size of the trees). The high accuracy of the automated labeling justifies expanding the analysis to a larger corpus of dialogues. The results presented below correspond to the automatically labeled corpus of 1676 dialogues (11,024 dialogue turns) that occurred during the months of March and April of 2008.

III. RELATING DISCOURSE FEATURES TO THE INITIATOR AND THE GREETING

A. Data analysis

In the following data analysis, we estimate the relation between two (not mutually exclusive) ways to begin a dialogue and the discourse: (1) whether the dialogue has been initiated by the robot and (2) whether the user has started the dialogue with a *Greeting* (e.g. "Hi", "Good morning"). We define a dialogue as initiated by the robot if the user started typing within 10 seconds from the time the robot has greeted a passer-by.

The features of dialogues that we compare include start time, dialogue duration in seconds, dialogue duration in number of turns, total number of user's words, average number of user's words per user's turn, user's *Goodbye* as their last turn (*EndingWithBye*), robot's admitting its failure to make sense of the preceding user's turn (*InterpNegFeedback*), user's rude language (*Rude*), user's *Persistence*—robot's *InterpNegFeedback* followed by a non-empty user's turn that is not a *Goodbye*, *UserQuestion*, user's *QuestionAnswered* (i.e. question was parsed correctly and received a reasonable answer[2]), and user thanking the robot after the question has been reasonably answered (*AnswerThanked*). Under user greeting/no-greeting conditions we also compare the total number of user's words and average number of user's words per user's turn for the "inner" dialogue turns that exclude the two user turns trivially affected by the presence of a *Greeting*: an initial *Greeting* and trailing *Goodbye*.

The results are shown in Tables I and II. Where the units are not specified, the number represents the fraction of all the relevant dialogues where the respective dialogue pattern is present. For example, the fraction of dialogues containing the *QuestionAnswered* pattern is counted only among the dialogues that include user's questions, and the fraction of dialogues containing the *AnswerThanked* pattern is counted only among the dialogues that include user's questions that were reasonably answered by the robot. Differences of the means of variables that represent numerical counts or times are tested for significance by a two-sample t-test. The 2-by-2 contingency tables that show counts of dialogues containing respective discourse patterns among all the relevant dialogues that (a) were initiated by the robot/user, or (b) do/do not include greeting in the user's first turn, are tested for independence using Pearson's Chi-squared test with Yates' continuity correction.[3]

	robot-init.	user-init.	p-value
start time	2:29pm	2:43pm	0.16
duration (sec)	38.95	47.77	0.38
num. of all turns	6.34	6.70	0.15
num. of user's words	9.10	9.28	0.70
user's words per user's turn	2.70	2.70	0.99
Greeting	0.43	0.38	0.05
EndingWithBye	0.13**	0.18**	< 0.01
InterpNegFeedback	0.51	0.53	0.40
Rude	0.03	0.02	0.38
Persistence	0.72	0.70	0.62
UserQuestion	0.58	0.62	0.19
QuestionAnswered	0.50**	0.40**	< 0.01
AnswerThanked	0.15	0.17	0.63

Table I. Associations between the initiator of the dialogue and the discourse, using dialogue turns labeled by a classifier. Values marked with $*$ and $**$ correspond to significant results at 0.05 and 0.01 levels respectively.

B. Discussion

While the initiator of the dialogue does not show as much effect on the discourse as whether the user started with a *Greeting*, robot-initiated dialogues show a slight increase in the fraction of dialogues with *QuestionAnswered*, user's *Greeting*, and a negative effect on *EndingWithBye*. Further analysis is necessary to explain these differences.

The effect of the user's *Greeting* on the length of the dialogue in terms of the number of turns can be explained

[2] We are loose in our interpretation of what constitutes a reasonable answer. We consider an answer to any plausible interpretation of the question as reasonable. For example, the robot's answer in the second line of Figure 2 is reasonable, albeit unhelpful.

[3] An abridged version of this corpus analysis has been presented as a short paper at HRI'09 [12].

	Greeting	¬*Greeting*	p-value
start time	2:24pm*	2:47pm*	0.02
duration (sec)	43.61	45.44	0.84
num. of all turns	7.66**	5.88**	< 0.01
num. of user's words	10.16**	8.60**	< 0.01
user's words per user's turn	2.34**	2.94**	< 0.01
num. of user's words (inner)	8.35	8.31	0.92
user's words per u. turn (inner)	3.18*	3.00*	0.03
num. of user's turns (inner)	2.61	2.80	0.12
EndingWithBye	0.21**	0.14**	< 0.01
InterpNegFeedback	0.46**	0.57**	< 0.01
Rude	0.03	0.02	0.62
Persistence	0.76**	0.67**	< 0.01
UserQuestion	0.61	0.60	0.66
QuestionAnswered	0.48*	0.41*	0.02
AnswerThanked	0.25**	0.09**	< 0.01

Table II. Associations between a *Greeting* in the user's first turn and the discourse, using dialogue turns labeled by a classifier. Values marked with $*$ and $**$ correspond to significant results at 0.05 and 0.01 levels respectively.

by the additional pair of greeting turns. It appears that presence of a *Greeting* does not change the overall verbosity of the dialogue when the *Greeting* and *Goodbye* turns are excluded. However, the average length (words per turn) of the inner turns is slightly larger for the interactions that start with a *Greeting*. Users starting with a *Greeting* tend to exhibit more *Persistence* and are more than 2.6 times are as likely to thank the robot after it answers their question (*AnswerThanked*). They also have a better chance of having their questions parsed. The fact that users that start with a *Greeting* have fewer chances of being not understood by the robot (*InterpNegFeedback*) is not explained by the presence of trivial "hi-hi-bye-bye" interactions, since the difference remains significant for interactions containing more than 4 turns (not shown in the Table).

IV. ANALYSIS OF USER'S THANKING BEHAVIOR

The analysis presented in Section III uncovered the tendency of users who greet the robot to also say thanks after having their questions answered. In this section, we explore the user's thanking behavior in more detail.

The annotation scheme we have used so far does not differentiate between the types of the questions and is loose in defining what it means to have the question reasonably answered. For example, in our particular context, the question "How old are you?" does not warrant "thanks" even after a helpful answer. On the other hand, an answer to another information seeking question, "Where is Wean Hall?" can be followed by "Thanks." We hypothesize a relationship between the relevance, or helpfulness, of the robot's answer and the presence of thanks in the following user's turn. For example, if the robot misinterpreted the question and gave directions to a location that is different from the one intended by the user, the answer, while reasonable, is certainly not helpful (as in the example in Figure 2). In this section, we use a finer grained manual annotation of the user's questions and robot's answers to uncover the relationship between the presence of user's greeting, helpfulness of the robot's answer, and whether the user thanked the robot. In particular, we describe and evaluate a classifier of *Unhelpful* robot's answers that uses the presence of *Thanks* in the user's following turn and the presence of *Greeting* in the user's first turn as its features.

A. Manually labeled corpus of thanking behavior

We use a finer grained annotation for the analysis of thanking behavior that consists of labeling user's turns as *Thankable* and *Non-thankable* questions and robot's responses as *Helpful* and *Unhelpful answers*. We extended our manual annotation of the 8-day corpus of 287 dialogues with these additional labels, considering partial answers as unhelpful and excluding the combinations of an answer and a followup question, e.g. "The Robotics Education Lab is in NSH 3206. Would you like directions?", where the thanking behavior is complicated by the interference with an answer to the question "Would you like directions?".

Notice, that the label *Thankable* question is defined by semantics and pragmatics, rather than by syntactic features, so it is possible that an utterance is a *Thankable* question, but not a *UserQuestion* according to the previous, more syntactically-biased labeling scheme, as in "I meant another Frank." (Figure 2). The contingency table of *UserQuestions* and *Thankable* questions among all user utterances is presented in Table III.

	Thankable question	¬*Thankable* question
UserQuestion	148	223
¬*UserQuestion*	58	549

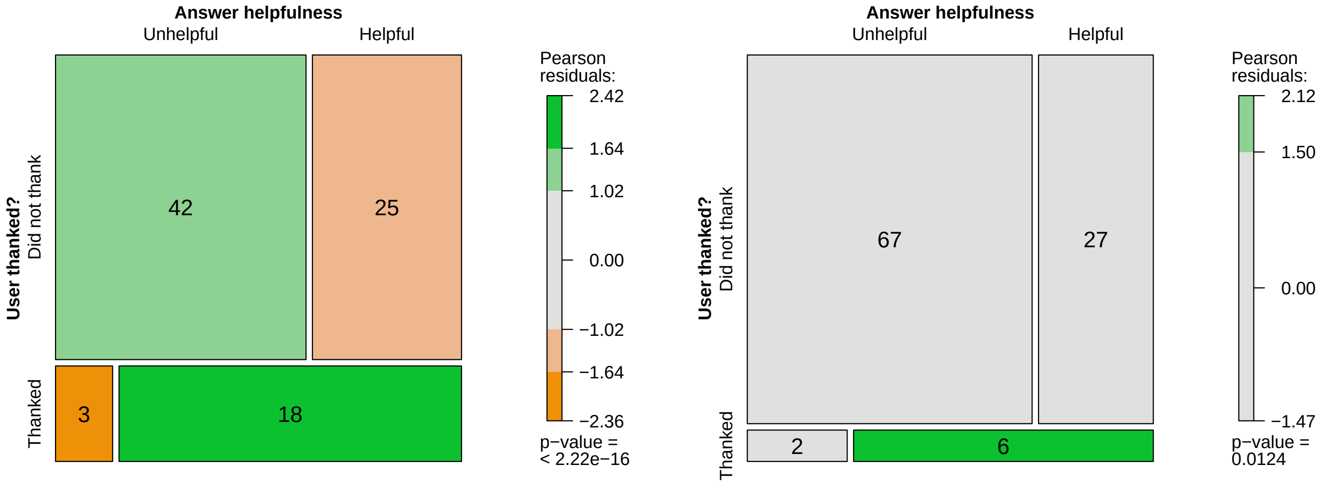
Table III. Distribution of *UserQuestions* and *Thankable* questions among all user turns.

The split of the *Thankable* questions between users with and without *Greeting* in their first turn is 92 to 114 (no significant departure from the independence hypothesis according to Pearson's Chi-squared test). 16 of these *Thankable* questions are followed by a combination of an answer and a followup question and are removed from further analysis.

B. Data analysis

We restrict our following analysis to the 190 sequences of dialogue turns that contain a *Thankable* question, e.g. <*Thankable* question, *Unhelpful* answer, *Thanks*>. A single dialogue can contain multiple sequences. Figures 3 and 4 show the mosaic plots of the counts of discourse patterns starting with the user's *Thankable* question, followed by the robot's *Helpful*, or *Unhelpful*, answer that were or were not, followed by the user's *Thanks* for three sets of dialogues: (a) dialogues where users greeted the robot on their first turn, (b) dialogues where users did not greet the robot on their first turn, and (c) all dialogues.

Color shading indicates cells responsible for the Pearson residuals that exceed (in absolute value) critical values corresponding to 0.1 and 0.01 levels. The departure from independence between user's *Thanks* and the helpfulness of the answer is significant ($p < 0.01$), both conditionally on the user's *Greeting* and for all users pooled together.



(a) Users who greeted the robot.

(b) Users who did not greet the robot.

Fig. 3. Thanking after a helpful answer within (a) users who greeted and (b) users who did not greet the robot. The shading is based on the maximum absolute values of Pearson residuals statistic. Cells shaded in light and in fully saturated colors correspond to residuals that exceed critical values of the permutation test for independence (conditional on *Greeting*) at 0.1 and 0.01 levels respectively.

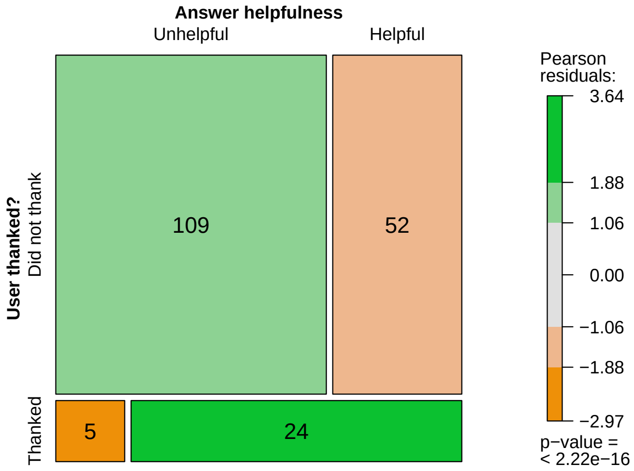


Fig. 4. Thanking after a helpful answer for all users. The shading is explained in the caption to Figure 3.

Conditionally on *Greeting*, the violation of the independence hypothesis is most prominent among the users who greeted the robot ($p < 0.1$ for all cells), especially when the users also *Thanked* the robot ($p < 0.01$). While users that did not greet also tend to thank less, both users that greeted and users that did not appear to thank a considerable fraction of unhelpful answers, in a sense of “Thanks anyways.”

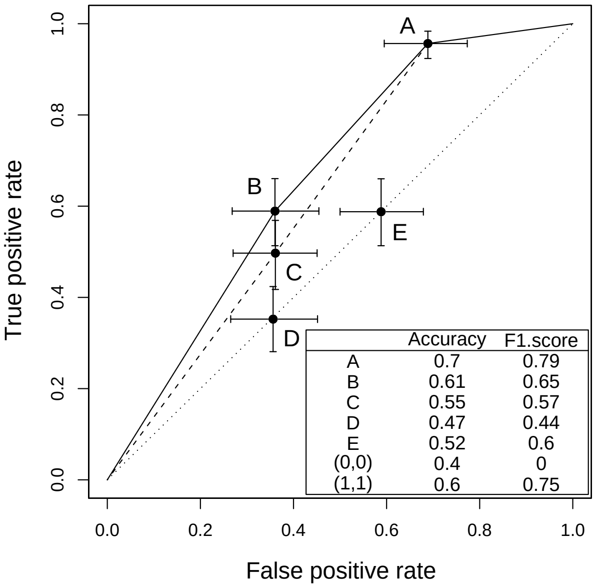
V. Detecting unhelpful answers

We use two binary features (*Greeting* in the first turn and *Thanks* in the turn following a robot’s answer) to predict a binary variable (*Unhelpful answer*), hence we can represent all possible deterministic classifiers by 16 decision trees. Two of the classifiers are trivial and always output *Helpful* or *Unhelpful*. We compare the performances of the set of two non-trivial classifiers that use *Thanks* as their only feature with the set of all 16 classifiers (12 of which use both features non-trivially) that use both features by comparing their ROC curves, namely the graphs of their respective true positive rates $tpr = tp/(tp + fn)$ versus false positive rates $fpr = fp/(fp + tn)$, where tp corresponds to the answers correctly detected *Unhelpful*, fp to the answers incorrectly detected as *Unhelpful*, tn to the answers correctly detected as *Helpful*, and fn to the answers incorrectly detected as *Helpful*. One of the properties of ROC space is that one could always combine a number of classifiers by random sampling in a way that the ROC points of their combinations would trace a convex hull of the ROC points/curves of the individual classifiers [11]. Therefore, to compare two sets of classifiers we have to consider only the classifiers that are on the convex hull of each of the sets. Since we can always add the two trivial constant classifiers to any classifier set, the convex hulls will always contain points $(0, 0)$ and $(1, 1)$. For example, the ROC curve corresponding to random combinations between these two trivial classifiers (i.e. with probability p ask classifier $(0, 0)$ that outputs *Helpful*, otherwise ask classifier $(1, 1)$ that outputs *Unhelpful*) is the dotted diagonal in Figure 5.

The dashed line in Figure 5 corresponds to the convex hull of the set of two classifiers that use *Thanks* as their only feature. From this set, only the classifier

A: *Thanked* → *Helpful*, else *Unhelpful*

is on the hull. The solid line corresponds to the convex hull of all 16 classifiers. The two classifiers that are on the latter convex hull are **A**, and the classifier that extends the condition in **A** to also predict *Helpful* for all users that greeted:



	Accuracy	F1.score
A	0.7	0.79
B	0.61	0.65
C	0.55	0.57
D	0.47	0.44
E	0.52	0.6
(0,0)	0.4	0
(1,1)	0.6	0.75

Fig. 5. True positive and false positive rates (ROC curves) for the classifiers of *Unhelpful answers* relying on the user's *Thanks* as the only feature (dashed line) and incorporating both the user's *Thanks* and *Greetings* as features (solid line). The detailed explanation is in the text.

B: *Thanked* $\vee$ *Greeted* $\rightarrow$ *Helpful*, else *Unhelpful.*

For comparison we also plot the performance of classifiers

C that randomly samples between **A** and the trivial "always output *Helpful*" classifier with a probability p such that its fpr is approximately equal to fpr of **B**, and

D and **E** that use no features and approximately match **B**'s fpr and tpr respectively.

The bars shown in Figure 5 correspond to the 90% empirical confidence intervals that are constructed by applying the bootstrap to generate 1000 samples from the original sample of 190 dialogue turn sequences (see [11] for details on using the bootstrap for ROC curve analysis). The accuracy and F1-scores shown in the figure are indicative of a class skew. Indeed, the slope of iso-accuracy lines is

$$\frac{fn+tn}{tp+fp} = \frac{\#\mathit{Helpful}}{\#\mathit{Unhelpful}} = \frac{76}{114} \approx 0.67.$$

While neither the difference in tpr between the two-feature classifier **B** and the single-feature classifier **C** nor the difference in tpr between **C** and the random classifier **D** is significant at 0.1 level, the improvements in tpr between **B** and **D**, and in fpr between **B** and **E** are both significant at 0.05 level.

VI. CONCLUSIONS

The analysis of the unconstrained human-robot dialogues that we presented has shown that the user's social dialogue acts, such as greeting and thanks are significantly associated with the certain types of system's errors and that greeting is associated with discourse patterns at various stages of the dialogue. In particular, the user's greeting in the first dialogue turn is associated with remote and global discourse features such as ending the dialogue with goodbye, persistence, and the likelihood of receiving an answer. User's thanks, while also associated with the greeting, is a considerable predictor of (un)helpful answers, especially when combined with the greeting. The seemingly global scope of associations involving the greeting in the user's first turn motivated us to treat this feature as a component of a user model.

Future work includes improving unhelpful answer detection by using additional lexical-level features of the following user turn [7], and by expanding the user model to include other features of the early dialogue turns.

Our larger goal, however, is to use the unhelpful answer detection and the user model to guide the dialogue. A robot, for example, could adapt its dialogue by providing additional encouragement to an anticipated non-persistent user. Detecting and adapting to its own unhelpfulness may give the robot a degree of meta-cognition that could improve the interaction experience.

REFERENCES

[1] H. Bunt. Dimensions in dialogue act annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, 2006.

[2] H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive Science*, 13:259–294, 1989.

[3] K. Forbes-Riley and D. Litman. Adapting to student uncertainty improves tutoring dialogues. In *Proc. 14th Int. Conf. on Artificial Intelligence in Education (AIED)*, Brighton, UK, July 2009.

[4] K. Forbes-Riley and D. Litman. Designing and evaluating an uncertainty-adaptive spoken dialogue tutoring systems. (under review). 2009.

[5] R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A. C. Schultz, , and J. Wang. Designing robots for long-term social interaction. In *Proc. Int. Conf. on Intelligent Robots and Systems*, pages 2199–2204, August 2005.

[6] J. Hirasawa, N. Miyazaki, M. Nakano, and K. Aikawa. New feature parameters for detecting misunderstandings in a spoken dialogue system. In *Proc. Int Conf. of Spoken Language Processing (ICSLP)*, volume 2, pages 154–157, Bejing, China, 2000.

[7] E. Krahmer, M. Swerts, M. Theune, and M. Weegels. Problem spotting in human-machine interaction. In *Proc. Eurospeech*, volume 3, pages 1423–1426, 1999.

[8] E. Krahmer, M. Swerts, M. Theune, and M. Weegels. Error detection in spoken human machine interaction. *Int. J. of Speech Technology*, 4(1):19–29, 2001.

[9] M. K. Lee and M. Makatchev. How do people talk with a robot? An analysis of human-robot dialogues in the real world. In *Proceedings of CHI*, pages 3769–3774, April 2009.

[10] D. J. Litman, J. B. Hirschberg, and M. Swerts. Predicting automatic speech recognition performance using prosodic cues. In *Proc. NAACL*, pages 218–225, 2000.

[11] S. A. Macskassy, F. Provost, and M. L. Littman. Confidence bands for ROC curves. In *CeDER Working Paper, IS-03-04*, New York University, NY, 2004. Stern School of Business.

[12] M. Makatchev, M. K. Lee, and R. Simmons. Relating initial turns of human-robot dialogues to discourse. In *Proceedings of the International Conference on Human-Robot Interaction (HRI)*, March 2009.

[13] T. Paek. Toward a taxonomy of communication errors. In *Proc. ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, pages 53–58, 2003.

[14] A. Roque and D. Traum. Improving a virtual human using a model of degrees of grounding. In *Proc. IJCAI*, 2009.

[15] M. Swerts, D. Litman, and J. Hirschberg. Corrections in spoken dialogue systems. In *Proc. ICSLP-2000*, pages 615–618, 2000.

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