A User Modeling-based Performance Analysis of a Wizarded Uncertainty-Adaptive Dialogue System Corpus

Kate Forbes-Riley, Diane Litman

Learning Research and Development Center (LRDC), University of Pittsburgh, USA
forbesk@cs.pitt.edu, litman@cs.pitt.edu

Abstract

Motivated by prior spoken dialogue system research in user modeling, we analyze interactions between performance and user class in a dataset previously collected with two wizarded spoken dialogue tutoring systems that adapt to user uncertainty. We focus on user classes defined by expertise level and gender, and on both objective (learning) and subjective (user satisfaction) performance metrics. We find that lower expertise users learn best from one adaptive system but prefer the other, while higher expertise users learned more from one adaptive system but didn't prefer either. Female users both learn best from and prefer the same adaptive system, while males preferred one adaptive system but didn't learn more from either. Our results yield an empirical basis for future investigations into whether adaptive system performance can improve by adapting to user uncertainty differently based on user class.

Index Terms: user modeling, affect/attitude adaptation, spoken dialogue, tutoring system, subjective and objective metrics

1. Introduction

There is increasing interest in building dialogue systems to detect and adapt to user affect, attitude and other metacognitive states. Promising results have been reported on automatic detection of such user states (e.g., [1]). A few experiments have further shown that by detecting and adapting to such user states, system performance (e.g., as measured by user satisfaction [2] and student learning [3]) can be improved compared to non-adaptive baseline systems.

However, while this line of research is promising, there is still much room for improvement. We are exploring user modeling techniques as one way of potentially improving the effectiveness of our adaptive tutoring system that adapts to user uncertainty. Prior user modeling research has shown that not all users interact with dialogue systems in the same way [4]. Research has further shown that developing and implementing different system behaviors for different user classes can improve performance [5, 6]. This user modeling approach can be divided into two main areas. Dynamic approaches determine a user's class based on features collected during system use (e.g., confidence scores from the speech recognizer [5]). Static approaches determine a user's class based on user features obtained before runtime (e.g., a user's age [7]). Some user classes are determined by both dynamic and static approaches (e.g., domain expertise [8]). In some domains (e.g., many call-center systems) it can be difficult under either a static or dynamic approach to obtain enough information about the user to accurately determine to which class a user belongs [9]. However, other domains, such as medical or tutoring, employ instruments such as surveys and tests to permit the extraction of a wealth of static user class information, including age, gender, expertise, etc. These domains may also acquire repeat users, yielding additional class-based information derived from users' past system interactions that can be used in either a static or dynamic approach.

To our knowledge, little (if any) prior spoken dialogue system research has applied user modeling techniques to the problem of adaptation to user metacognitive states. Here we analyze interactions between performance and user class in a dataset previously collected with two wizarded spoken dialogue tutoring systems that adapt to user uncertainty. We take a static approach, focusing on user classes defined by expertise level and gender, which are easily obtained in our system. To evaluate performance, we consider both objective and subjective metrics: student learning and user satisfaction. Objective and subjective metrics both represent crucial aspects of performance in most types of spoken dialogue systems [10]. For example, in call center systems, task completion (objective) and ease of use (subjective) are both important, as a user won't use a system if it is too difficult, even if they can complete their task with it. Similarly, in tutoring systems, student learning and user satisfaction are both important, as a student won't use a system if s/he doesn't like it, even if they can learn from it.

Our user modeling analysis indicates for whom uncertainty adaptations are effective and in what way. In particular, lower expertise users learn more from one adaptive system but prefer the other, while higher expertise users learned more from one adaptive system but didn't prefer either. Female users learn more from and prefer the same adaptive system, while males preferred one adaptive system but didn't learn more from either. These results motivate a future system redesign to investigate if spoken dialogue tutoring system performance can be improved by adapting to uncertainty differently based on user class.

2. System and Data

ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) is built on top of the Why2-Atlas text-based tutor [11]. ITSPOKE tutors 5 qualitative physics problems over 5 dialogues, in a question - answer - response format. In the original (non-adaptive) ITSPOKE, tutor responses depend only on the correctness of student answers. In our two adaptive ITSPOKEs (*basic* and *empirical*), tutor responses depend on both the correctness and the uncertainty of student answers. These adaptations, and the experiment in which we evaluated them, are summarized below and discussed in detail elsewhere [3].

In *basic* adaptive ITSPOKE, tutor responses are determined as follows: If the student answers *correctly without uncertainty*, ITSPOKE responds with Correctness feedback (e.g., "Right"). If the student answers *incorrectly with or without uncertainty*, ITSPOKE responds with Incorrectness feedback (e.g., "Well...") and additional content that walks the student

through the correct line of reasoning. Finally, if the student answers *correctly with uncertainty*, ITSPOKE gives the same response that it would give if the answer were incorrect (except with Correctness feedback, e.g., "Fine.").

T1: Now let's talk about the net force exerted on the truck. By the same reasoning that we used for the car, what's the overall net force on the truck equal to?
S1: The force of the car on it?? *[CU]*
T2: Fine. We can derive the net force on the truck by summing the individual forces on it, like we did for the car. First, what horizontal force is exerted on the truck during the collision?

Figure 1: Example of *Basic* Uncertainty Adaptation

Our *basic* adaptation is illustrated in Figure 1. It derives from tutoring theory that views both uncertainty and incorrectness as signals of "learning impasses" (e.g., [12]). We distinguished three impasse types by combining binary uncertainty (uncertain(U), nonuncertain (nonU)[1]) and correctness (incorrect (I), correct (C)): **InonU, IU, CU**. Our *basic* adaptation provided the same additional content (the incorrect answer response) to remediate *all* impasses (CU, IU, InonU). In contrast, the original ITSPOKE remediated only incorrectness impasses (IU, InonU) and so ignored one uncertainty impasse (CU). Both *basic* and original ITSPOKE gave feedback that varied based only on the answer's correctness (ignoring uncertainty).

Our *empirical* adaptation revised our *basic* adaptation based on empirical analyses of human tutor responses; it provided additional content to remediate all impasses (CU, IU, InonU), but varied both *the dialogue act* used to present this content and *the feedback*, based on the impasse type. One example is shown in Figure 2. The feedback variations acknowledge the (in)correctness content and the uncertainty content, and were based on prior research showing that human tutor-derived empathetic system responses can positively impact performance (e.g., [13]). The dialogue act variations were based on dialogue acts we found a human tutor to use significantly more or less than expected after each impasse type. After CUs we used a "Bottom Out" version of the original incorrect answer response. This is illustrated in Figure 2 (compare with the original response in **T2** in Figure 1). After IUs, we used a short Bottom Out followed by one or more Short Answer Questions, while after InonUs we used one or more Short Answer Questions.

S1: The car's force hitting the truck?? *[CU]*
T2: That's exactly right, but you seem unsure, so let's sum up. The net force on the truck is equal to the impact force on it. We can prove this just like we did for the car. First, we know gravity and the normal force on the truck must cancel each other, otherwise the truck would not be at rest vertically. Second we know that the impact force is the only horizontal force exerted on the truck.

Figure 2: Example of *Empirical* Uncertainty Adaptation

Our experiment used a Wizard of Oz scenario: a human "wizard" performed speech recognition, language understanding, and uncertainty annotation. One control condition (*normal*) used our original ITSPOKE. A second control condition (*random*) also used this ITSPOKE but treated a percentage of random correct answers as incorrect, to control for the additional content in the experimental conditions. The first experimental condition (*basic*) used *basic* ITSPOKE. The second experimental condition (*empirical*) used *empirical* ITSPOKE. Subjects: read a short physics text; took a pretest; worked 5 problems with ITSPOKE; took a survey (Figure 3); took a posttest.

[1] A 'nonuncertain' answer may be certain or neutral for certainty.

3. User Classes Performance Analysis

3.1. Method

We previously examined the main effect of condition on learning and user satisfaction metrics over all users (see [3]). With respect to the metrics in this paper (see below), we find a main effect of learning gain and pairwise tests show users learned more from *basic* than *norm* or *empirical*, but average user posttest score with *basic* was only 81% (% correct). Moreover, although there are no main effects for our three user satisfaction metrics, pairwise tests show users preferred *empirical* to *basic* with respect to the quality of the spoken dialogue interaction. These results suggest that neither adaptive system is maximally effective. A maximally effective system would yield strong performance on both objective (e.g., learning gain) and subjective (e.g., user satisfaction) evaluation metrics.

We hypothesized that the inconsistency in our main effects results might be due to different classes of users in our corpus. That is, some users might learn more and/or prefer *basic*, others might learn more and/or prefer *empirical*, while others might not benefit from either adaptation, but would benefit from a different uncertainty adaptation. If we could identify these users, we could potentially redesign a more effective system that adapted differently to uncertainty for each user class.

As discussed in Section 1, user modeling techniques provide a method of exploring this hypothesis. Here we analyze different user classes in the corpus collected in our prior experiment (Section 2). Our user classes are based on information that users supplied before interacting with our tutoring system, and are applicable to all types of spoken dialogue systems.

First, we hypothesized that users with different levels of "domain expertise" might benefit differently from our two uncertainty-adaptive systems. User expertise has been used in prior static user modeling approaches (e.g., [8]) and has been shown to be relevant to learning in computer tutoring (e.g., [14]). In our study, users with pretest scores below the mean (over all users) were put in the class *lower*, all others were put in *higher*. A t-test showed the higher and lower classes represent different populations ($p < 0.001$), and there was no significant difference in pretest score across conditions (within classes or overall). Note that the *higher* class was not expert in the physics domain; only physics novices were solicited for the experiment, and average pretest scores were 64% and 40% across conditions for the higher and lower classes, respectively. These averages indicate both expertise classes could benefit substantially from the tutoring in terms of learning.

Second, we hypothesized that different genders might benefit differently from our two uncertainty-adaptive systems. Prior studies have investigated whether genders behave differently with dialogue systems (e.g., [15]) and human tutors (e.g., [16]); for example, [16] find that tutorial dialogue structure is influenced by student expertise, gender, and self-efficacy levels.

For each binary category (expertise and gender), we investigated whether the two user classes patterned differently on both objective and subjective evaluation metrics.[2] As discussed in Section 1, both types of metrics are important in spoken dialogue systems in general, and in tutoring systems in particular. We used normalized learning gain as our objective metric ((posttest-pretest)/(1-pretest)). We used three subjective user satisfaction metrics, each representing a specific type of preference formed by totaling the user ratings for a specific group of

[2] Data sparsity prevented statistical analysis of combined classes (e.g. lower expertise males).

questions in our user satisfaction survey in Figure 3. Questions 1-7 are taken from [17] and 8-9 were created for our system; these questions concern the tutoring domain. Questions 10-12 were created for our system and concern the uncertainty adaptations. Questions 13-16 are taken from [18] and concern the spoken dialogue interaction.

Q1: It was easy to learn from the tutor.
Q2: The tutor didn't interfere with my understanding of the content.
Q3: The tutor believed I was knowledgeable.
Q4: The tutor was useful.
Q5: The tutor was effective on conveying ideas.
Q6: The tutor was precise in providing advice.
Q7: The tutor helped me to concentrate.
Q8: The tutor responded effectively after I was incorrect about the answer to a question.
Q9: The tutor responded effectively after I was correct about the answer to a question.

Q10: The tutor responded effectively after I was uncertain about the answer to a question.
Q11: The tutor responded effectively after I was certain about the answer to a question.
Q12: The tutor's responses decreased my uncertainty about my understanding of the content.

Q13: It was easy to understand the tutor speech.
Q14: I knew what I could say or do at each point in the conversations with the tutor.
Q15: The tutor worked the way I expected it to.
Q16: Based on my experience using the tutor to learn physics, I would like to use such a tutor regularly.

ALMOST ALWAYS (5), OFTEN (4), SOMETIMES (3), RARELY (2), ALMOST NEVER (1)

Figure 3: ITSPOKE Survey

For each metric, we ran a 2x4 factorial ANOVA with pairwise simple effects tests on the binary user class factor and the quaternary condition factor, to analyze the interaction effect of condition and user class and determine if the two adaptive systems perform differently based on class. If so, this suggests that performance can be improved by redesigning our system to adapt differently to uncertainty for the two classes.

3.2. Results

Table 1: Performance Results for Lower Expertise Users

Metric	Cond(N)	Mean	Diff	p
Learning Gain	*norm* (10)	0.375	-	
	rand (12)	0.538	-	
	basic (11)	0.619	$>$ *norm*	0.018
	empir (10)	0.448	-	
Spoken Dialogue Q13-Q16	*norm*	14.10	-	
	rand	15.25	-	
	basic	13.91	$<$ *empir*	0.023
	empir	16.30	$>$ *norm*	0.041

Table 2: Performance Results for Higher Expertise Users

Metric	Cond (N)	Mean	Diff	p
Learning Gain	*norm* (11)	0.389	-	
	rand (8)	0.563	-	
	basic (9)	0.593	$>$ *empir*	0.039
	empir (10)	0.369	-	

For our user expertise classes, no metric showed a significant overall interaction effect in the initial ANOVA, but multiple metrics showed significant interactions in the pairwise tests. These tests compared each pairwise combination of condition and class, to determine if they were significantly different for the metric. We consider these pairwise test results most useful from a system redesign perspective: for each class, they tell us specifically what's working or not with each adaptive system with respect to each other and the non-adaptive systems.

Tables 1-2 show metrics yielding significant ($p < 0.050$) results on the pairwise tests for the expertise classes. The columns show the metric, the condition (and number of users), its mean, the condition with which a difference is found, and the direction ($>$ or $<$) and significance of this difference. As we expected, the expertise classes pattern differently. Table 1 shows that lower expertise users learn significantly more from *basic* than *normal*, but prefer *empirical* over both *basic* and *normal* with respect to spoken dialogue interaction quality. We hypothesize that lower expertise users perceived the spoken dialogue in *empirical* as easier to follow because, unlike the other systems, the feedback in *empirical* always responded explicitly to the uncertainty of their answers, as well as the correctness. Therefore, we hypothesize that lower expertise users would show greater user satisfaction with the *basic* adaptation if we modified it to include this type of feedback. This hypothesis is supported by other tutoring system research (e.g. [13]) showing that affect-related feedback increases user satisfaction.

However, our investigation suggests that the benefit of such feedback may only hold for lower expertise users; Table 2 shows that higher expertise users express no preference for any system. Moreover, although the higher expertise users learn significantly more from *basic* than *empirical*, neither adaptive system outperforms the baseline non-adaptive systems for learning. Therefore, we hypothesize that the *basic* adaptation is a better choice than *empirical* for higher expertise users, but it needs modification to be more effective. We discuss this in Section 4.

Table 3: Performance Results for Female Users

Metric	Cond(N)	Mean	Diff	p
Learning Gain	*norm* (14)	0.375	-	
	rand (11)	0.516	-	
	basic (12)	0.597	$>$ *norm*	0.017
	empir (12)	0.401	$<$ *basic*	0.041
Uncertainty Adaptation Q10-Q12	*norm*	11.07	-	
	rand	12.27	-	
	basic	12.75	$>$ *norm*	0.020
	empir	11.33	-	

Table 4: Performance Results for Male Users

Metric	Cond(N)	Mean	Diff	p
Tutoring Q1-Q9	*norm* (7)	39.00	$>$ *basic*	0.038
	rand (9)	37.00	-	
	basic (8)	34.25	-	
	empir (8)	39.00	$>$ *basic*	0.032
Spoken Dialogue Q13-Q16	*norm*	16.00	$>$ *basic*	0.009
	rand	15.33	$>$ *basic*	0.027
	basic	12.88	-	
	empir	16.63	$>$ *basic*	0.001

For our gender classes, we found that two user satisfaction metrics showed a significant overall interaction effect in the initial ANOVA (Q10-12: $F(3,73) = 3.711$, $p=0.015$; Q13-16: $F(3,73) = 3.429$, $p=0.021$). Tables 3-4 show metrics yielding significant ($p < 0.050$) pairwise test results for the two gender classes. As we expected, the two genders pattern differently. Table 3 shows that female users learn significantly more from *basic* than either *normal* or *empirical*. Moreover, females prefer *basic* to *normal* with respect to the quality of the uncer-

tainty adaptation. These results suggest that the *basic* uncertainty adaptation is reasonably effective for females, and that system redesign effort is best focused on other user classes who show less performance improvement from an adaptive system.

In particular, Table 4 shows that males achieve no significant learning difference with any system. Males prefer *empirical* to *basic* with respect to quality of both the tutoring and the spoken dialogue interaction; however, neither adaptive system outperforms the baseline non-adaptive systems for these metrics. Therefore, we hypothesize that the *empirical* adaptation is a better choice than *basic* for male users, but it needs modification to be more effective. We discuss this in Section 4.

Overall, our results shed new light on our initial analysis of the main effects of learning and user satisfaction over all users (see Section 3.1), which showed that users learned more from *basic* but preferred *empirical*'s spoken dialogue interaction. In fact, these main effects are primarily explained by lower expertise users; the other user classes differ on at least one metric.

4. Conclusions and Current Directions

We showed that a user modeling analysis of two uncertainty-adaptive spoken dialogue tutoring systems can indicate for whom the adaptations are working and in what way. Our results suggest that a more effective spoken dialogue tutoring system should adapt differently to user uncertainty based on user class. Our results also suggested specific hypotheses about how to adapt to uncertainty based on user class. In particular, the fact that lower expertise users learned more from *basic* but preferred *empirical* for spoken dialogue interaction quality suggested that lower expertise users would prefer *basic* with feedback that responded to both uncertainty and correctness. The fact that females both learned more from and preferred *basic* suggested that the *basic* adaptation is reasonably effective for females. The fact that higher expertise users learned more from *basic* but didn't prefer any system indicates that *basic* is a better choice than *empirical* but futher research is needed to determine a more effective adaptation for higher expertise users. The fact that males preferred *empirical* but didn't learn more from any system indicates that *empirical* is a better choice than *basic* but futher research is needed to determine a more effective adaptation for male users. In future work we will explore methods such as reinforcement learning and correlations of human and system tutor responses with evaluation metrics, to identify tutor responses to user uncertainty that both improve learning and yield increased user satisfaction for higher expertise users and males. We have also recently completed a fully automated version of our uncertainty adaptation experiment (where the system performs all tasks) and will analyze that corpus for similar user modeling results under these more realistic conditions.

Finally, note that other computer tutoring research has shown similar discrepancies between the system that yields the most learning and the system that is best-liked (e.g., [19]); we hypothesize that a user modeling approach to system redesign offers the best chance for developing a maximally effective system that improves both learning and user satisfaction. Of course, it may not be possible to find a strategy for every user class that optimizes both metrics. In such cases a design choice can be made to compromise one goal for a specific user class.

5. Acknowledgments

NSF (#0631930) supports this work. We thank the ITSPOKE Group, Bob Hausmann and Hua Ai for helpful comments.

6. References

[1] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, March 2005.

[2] K. Liu and R. W. Picard, "Embedded empathy in continuous, interactive health assessment," in *CHI Workshop on HCI Challenges in Health Assessment*, 2005.

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

[4] I. Zukerman and D. Litman, "Natural language processing and user modeling: Synergies and limitations," *User Modeling and User-Adapted Interaction*, vol. 11, no. 1/2, pp. 129–158, 2001.

[5] D. Litman and S. Pan, "Designing and evaluating an adaptive spoken dialogue system," *User Modeling and User-Adapted Interaction*, vol. 12, no. 2/3, pp. 111–137, 2002.

[6] J. Chu-Carroll and J. S. Nickerson, "Evaluating automatic dialogue strategy adaptation for a spoken dialogue system," in *Proc. NAACL*, 2000, pp. 202–209.

[7] K. Georgila, M. Wolters, and J. Moore, "Simulating the behaviour of older versus younger users when interaction with spoken dialogue systems," in *Proceedings of ACL-08: HLT Short Papers (Companion Volume)*, Columbus, Ohio, June 2008, pp. 49–52.

[8] S. Janarthanam and O. Lemon, "User simulations for online adaptation and knowledge-alignment in troubleshooting dialogue systems," in *Workshops on the Semantics and Pragmatics of Dialogues (SEMdial)*, London, June 2008.

[9] E. Reiter, S. Sripada, and S. Williams, "Acquiring and using limited user models in nlg," in *Proceedings of 2003 European Natural Language Generation Workshop*, 2003, pp. 87–94.

[10] S. Moller, P. Smeele, H. Boland, and J. Krebber, "Evaluating spoken dialogue systems according to de-facto standards: A case study," *Computer Speech and Language*, vol. 21, pp. 26–53, 2007.

[11] K. VanLehn, P. W. Jordan, C. P. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, R. Srivastava, and R. Wilson, "The architecture of Why2-Atlas: A coach for qualitative physics essay writing," in *Proc. Intelligent Tutoring Systems*, 2002.

[12] K. VanLehn, S. Siler, and C. Murray, "Why do only some events cause learning during human tutoring?" *Cognition and Instruction*, vol. 21, no. 3, pp. 209–249, 2003.

[13] W. Tsukahara and N. Ward, "Responding to subtle, fleeting changes in the user's internal state," in *Proc. SIG-CHI on Human factors in computing systems*. Seattle, WA: ACM, 2001.

[14] C. P. Rose, J. D. Moore, K. Vanlehn, and D. Allbritton, "A comparative evaluation of socratic versus didactic tutoring," in *Proc. Conference of the Cognitive Science Society*, 2001, pp. 869–874.

[15] K. Bhatt, M. Evens, and S. Argamon, "Hedged responses and expressions of affect in human/human and human/computer tutorial interactions," in *Proceedings of Cognitive Science (CogSci)*, Chicago, USA, 2004, pp. 114–119.

[16] K. E. Boyer, M. A. Vouk, and J. C. Lester, "The influence of learner characteristics on task-oriented tutorial dialogue," in *Proceedings of AIED*, Los Angeles, CA, 2007.

[17] A. L. Baylor, J. Ryu, and E. Shen, "The effect of pedagogical agent voice and animation on learning, motivation, and perceived persona," in *Proc. ED-MEDIA*, Honolulu, Hawaii, June 2003.

[18] M. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, E. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, S. Roukos, G. Sanders, S. Seneff, and D. Stallard, "DARPA Communicator: Cross-system results for the 2001 evaluation," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado, USA, 2002, pp. 269–272.

[19] K. Moreno, B. Klettke, K. Nibbaragandla, and A. Graesser, "Perceived characteristics and pedagogical efficacy of animated conversational agents," in *Proceedings of the Intelligent Tutoring Systems Conference (ITS)*, Biarritz, France, 2002.