Uncertainty Corpus: Resource to Study User Affect in Complex Spoken Dialogue Systems

Kate Forbes-Riley, Diane Litman, Scott Silliman, Amruta Purandare

University of Pittsburgh
Pittsburgh, PA 15260, USA
forbesk@pitt.edu, litman@cs.pitt.edu, scotts@cs.pitt.edu, amruta@cs.pitt.edu

Abstract

We present a corpus of spoken dialogues between students and an adaptive Wizard-of-Oz tutoring system, in which student uncertainty was manually annotated in real-time. We detail the corpus contents, including speech files, transcripts, annotations, and log files, and we discuss possible future uses by the computational linguistics community as a novel resource for studying naturally occurring user affect and adaptation in complex spoken dialogue systems.

1. Introduction

Within research on spoken dialogue systems, many promising results have been reported for automatically detecting user affective states (e.g., (Litman and Forbes-Riley, 2006; Vidrascu and Devillers, 2005; Batliner et al., 2003; Shafran et al., 2003)). The larger goal of this work is to improve spoken dialogue system performance by automatically adapting to user affect. The achievement of this goal could be significantly aided by studying affect-annotated corpora between users and spoken dialogue systems. However, to date only a few such corpora have been reported or made publicly available to the computational linguistics community. For example, while the HUMAINE project[1] contains a large collection of publicly available emotional speech corpora, very few contain naturally occurring human-computer dialogues (e.g. (Batliner et al., 2004; Walker et al., 2001; Ang et al., 2002)). Moreover, only the DARPA Communicator corpus uses English; it contains dialogues in the travel-planning (i.e. form-filling) domain, and user turns are annotated for frustration and annoyance.

To support further research towards the development of effective affect-adaptive systems, this paper presents another affect-annotated spoken dialogue system corpus, which uses English and reflects a complex human-computer interaction domain and new affect annotation. This *Uncertainty Corpus* contains spoken dialogues between students and a Wizard-of-Oz spoken dialogue tutoring system. The corpus was collected in a controlled experiment, in which student uncertainty was manually annotated in real-time by a human "Wizard", and was automatically adapted to in the experimental condition. This corpus is publicly available for scientific purposes (by request) through the Pittsburgh Science of Learning Center's Datashop[2]. We first describe the corpus collection. We then detail the corpus contents, including speech files, transcripts, annotations, and log files. Finally we discuss future uses by ourselves and the wider computational linguistics community as a novel resource for studying naturally occurring user affect and adaptation in complex (e.g. non-form filling) dialogue systems.

[1] http://emotion-research.net

[2] https://learnlab.web.cmu.edu/datashop/index.jsp

2. WOZ-TUT: Adaptive Wizard-of-Oz Spoken Dialogue Tutoring System

In prior work we developed ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) (Litman and Forbes-Riley, 2006). ITSPOKE tutors students in 5 qualitative physics problems. The dialogue manager uses a finite state paradigm; tutor responses (next states) are based on the correctness of the student answer (transitions between states).

We've begun enhancing ITSPOKE to respond to student affect[3] over and above correctness, and are initially targeting student uncertainty for two reasons. First, it occurred more often than other student affective states in our dialogues (Forbes-Riley and Litman, 2008). Second, although most tutoring systems respond based only on student (in)correctness, tutoring researchers view both incorrectness and uncertainty as signals of "learning impasses"; i.e. as opportunities for the student to engage in constructive learning (Craig et al., 2004; VanLehn et al., 2003). This view provides a straightforward adaptation hypothesis to test: Responding to student uncertainty in the same way as incorrectness should significantly increase learning, by providing students with knowledge to bridge their uncertainty impasses. Implementing this adaptation involved changing the next state transitions in the dialogue manager to depend on the answer's combined correctness and uncertainty value. That is, all uncertain+correct answers were treated as incorrect (uncertain+incorrect answers already are treated as incorrect).

We implemented this adaptation in a Wizard of Oz version of our ITSPOKE system that tutors only one physics problem. In this paper we will refer to this system as "WOZ-TUT". In WOZ-TUT, a few system components are re-

[3] We use "affect" to cover emotions and attitudes that can affect user communication in spoken dialogue. Some argue for separating the two, but some speech researchers find the narrow sense of "emotion" too restrictive since it excludes states in speech where emotion is present but not full-blown, including arousal and attitude (Cowie and Cornelius, 2003). Some tutoring researchers also combine emotion and attitude (e.g. (Pon-Barry et al., 2006; Bhatt et al., 2004)).

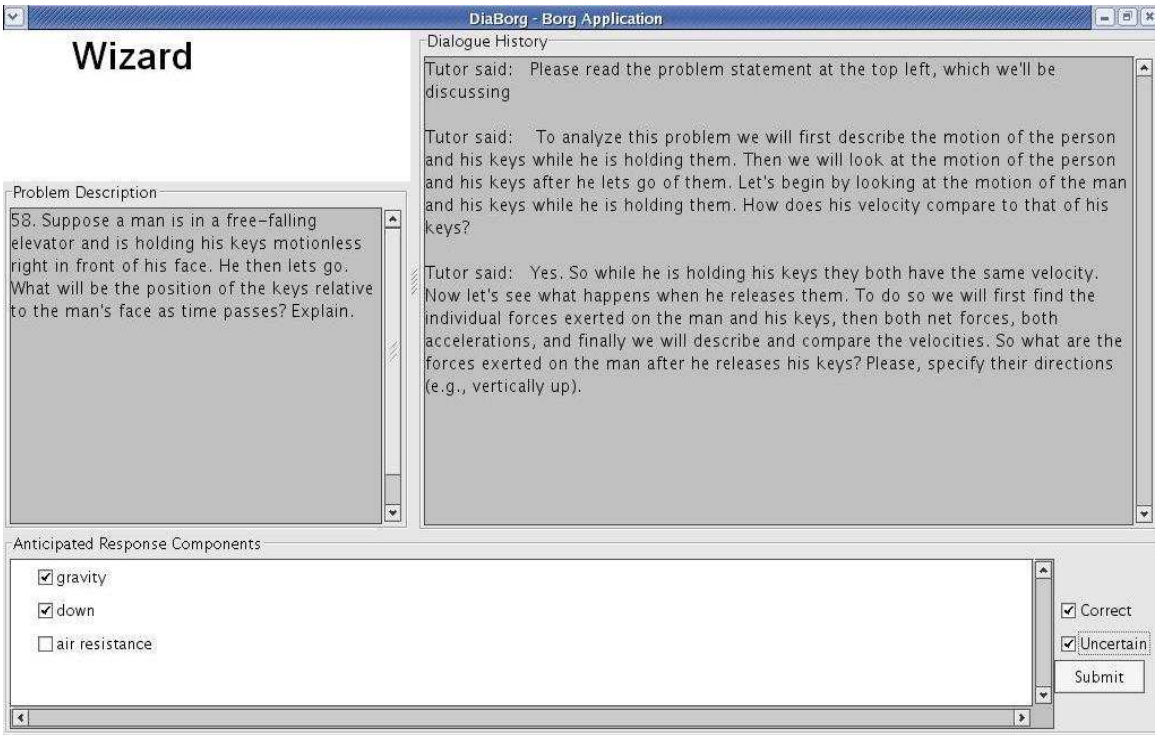

Figure 1: Screenshot of WOZ-TUT Wizard Interface

placed by a human "wizard". The wizard performs speech recognition, correctness annotation, and uncertainty annotation, for each student answer. In this way, we tested the upper bound performance of the uncertainty adaptation hypothesis without any potentially negative impact of automated versions of these three tasks.

Figure 1 shows a screenshot of the wizard's interface during the experiment. The physics problem is shown in the upper left box. A history of the text of the tutor turns is shown in the upper right box. The student turns aren't shown because they are spoken. Upon hearing each student answer, the Wizard annotates whether the answer is correct or uncertain in the lower right checkboxes.[4] Note that these correctness and uncertainty judgments are both binary. In other words, a "correct" answer may be either partially and fully correct, while a "nonuncertain" answer may be either certain or neutral for certainty (Forbes-Riley and Litman, 2008). These checkbox values are sent to the dialogue manager to determine the WOZ-TUT's response. In the lower left checkboxes, the Wizard annotates whether the answer that is heard is one that is "anticipated" by the system (specific responses to anticipated answers are authored in the system, while all other answers receive the same "unanticipated" response); these anticipated answers are logged for future analysis.

3. Experimental Design

The experiment had 3 conditions, designed to test whether our uncertainty adaptation would improve system performance (e.g. student learning). For use in these 3 conditions, the WOZ-TUT dialogue manager was parameterized, so that it could adapt contingently on the student state of uncertain+correct as discussed above, or randomly, or not at all.

In the experimental condition, the WOZ-TUT dialogue manager adapted to student uncertainty by treating all uncertain+correct student answers as incorrect. In the first control condition, the WOZ-TUT dialogue manager did not adapt to uncertainty (uncertainty was merely logged); it only treated incorrect answers as incorrect. In the second control condition, the WOZ-TUT dialogue manager did not respond to uncertainty (uncertainty was merely logged), but it did treat a percentage of random correct answers as incorrect, to control for the additional tutoring in the experimental condition. This percentage was toggled during the experiment to be similar to the percentage of uncertain+correct answers adapted to in the experimental condition.

Figures 2-3 illustrate how the WOZ-TUT responses differ across conditions. In both figures, **STUDENT**$_1$ is *uncertain+correct*. In Figure 2, **TUTOR**$_2$ asks an additional remediation question to remediate the student's uncertainty (experimental condition). In Figure 3, **TUTOR**$_2$ does not ask this remediation question (first control condition). An example from the second control condition is not shown because the treatment of any correct student answer will depend on whether it was randomly selected to receive the uncertainty adaptation. For example, in the second control condition, **STUDENT**$_1$ would be treated as shown in Figure 2 if it were randomly selected to receive the uncertainty adaptation, otherwise it would be treated as shown in Figure 3.

[4] In similar ITSPOKE corpora, this wizard displayed interannotator agreement of 0.85 Kappa on labeling binary correctness, and 0.62 Kappa on labeling binary uncertainty (Forbes-Riley and Litman, 2008).

TUTOR$_1$: What will the velocity of the object be a second after that (where the initial velocity is 9.8m/s and the acceleration is 9.8m/s^2)?

STUDENT$_1$: Uh. Nineteen point six meters per second?? *[uncertain+correct]*

TUTOR$_2$: Okay. As we have seen, if a falling object has an acceleration of 9.8m/s^2, that means that its velocity changes by 9.8m/s every second. So if a second after it began falling the object's velocity is 9.8m/s, a second later its velocity will be 9.8m/s+9.8m/s=19.6m/s. So what will its velocity be a second after it is 19.6m/s?

STUDENT$_2$: Um. I'm thinking. Twenty nine point four meters per second. *[nonuncertain+correct]*

TUTOR$_3$: Right. So at every point in time during the fall of the man and his keys, how do their velocities compare with each other?

Figure 2: Corpus Excerpt: Experimental Condition

TUTOR$_1$: What will the velocity of the object be a second after that (where the initial velocity is 9.8m/s and the acceleration is 9.8m/s^2)?

STUDENT$_1$: Nineteen point six meters per second?? *[uncertain+correct]*

TUTOR$_2$: Good. So at every point in time during the fall of the man and his keys, how do their velocities compare with each other?

Figure 3: Corresponding Corpus Excerpt: First Control Condition

The experimental procedure for collecting the Uncertainty Corpus was as follows. 60 subjects were randomly assigned to the 3 conditions, except that conditions were gender-balanced. The subjects were native English speakers who had not taken college-level physics. Each subject: 1) read background physics material; 2) completed a pretest; 3) used a web/voice interface to work through one physics problem with a version of WOZ-TUT (depending on condition); 4) completed a posttest; 5) worked a second problem isomorphic to the first problem with the non-adaptive WOZ-TUT (from the first control condition). We discuss uses of this isomorphic second problem in Section 5.[5]

4. Uncertainty Corpus Description

The resulting Uncertainty Corpus consists of 120 digitally recorded (.ogg format) dialogues from 60 students, totaling approximately 20 hours of dialogue. The tutor turn text sent to the text-to-speech synthesis was recorded in the log files, as were the correctness and uncertainty annotations of the student turns (labeled by the Wizard). The student turns were transcribed manually in separate files by professional transcribers after the experiment. These transcriptions include the turn text and endpoints, as well as punctuation and annotation of disfluencies and non-syntactic questions (the "??" in **STUDENT**$_1$, Figures 2-3). Transcription documentation is available with the corpus distribution. Table 1 provides further corpus details.

Table 2 shows differences in student answer attributes within problem and condition. Considering the first problem, one-way ANOVAs with pair-wise Tukey post-hoc analysis showed no significant difference between any of the three conditions in number of correct answers, uncertain answers, or uncertain+correct answers. A one-way ANOVA also showed no significant difference in the experimental (Exp) and second control (Ctrl2) conditions between the number of correct answers that received the adaptation. This confirms there was no experimenter bias. As shown in the last results row in the "Problem 1" section of Table 2, 36% of the random correct answers that received the adaptation in the second control (Ctrl2) condition were uncertain; thus 64% of these adapted-to answers were nonuncertain.

	Student	Tutor
Total Turns	2171	2531
Total Uncertain Turns	796	–
Total Words	13533	111829
Average Words/Turn	6.23	44.2

Table 1: Uncertainty Corpus Features

5. Uses of the Uncertainty Corpus

We see numerous uses of the Uncertainty corpus by ourselves and the wider computational linguistics community. Two potential uses are discussed below.

One use of the Uncertainty corpus is to compare system performance across conditions and isolate any impact of the affect adaptation. This use is relevant to researchers engaged in affect-adaptive spoken dialogue system development and evaluation. To date we have performed preliminary system performance analyses across conditions; further comparative analyses are on-going. These analyses fall into two main types. For our first analysis of system performance, we have performed statistical comparisons of student learning gains across conditions using our pretest

[5]The problem statement for this isomorphic problem was as follows: *A professor is sitting in his armchair holding his spectacles motionless in front of his face, in order to inspect how well he just cleaned them. All of a sudden, an earthquake causes his office floor to collapse. The professor is startled when he finds that both he and his armchair are in free-fall, and he drops his spectacles. What will be the position of the spectacles relative to the professor's face as time passes? Explain.*

	Exp	Ctrl1	Ctrl2
Problem1			
Ave. # Turns	20.65	18.60	19.75
Ave. # Correct Turns	13.80	12.55	14.20
Ave. # Uncertain Turns	9.95	8.60	11.15
Ave. # Uncertain+Correct Turns	4.75	3.75	6.10
Ave. # Correct Turns Adapted To (*treated as Incorrect*)	4.75	0	3.65
Ave. % Uncertain+Correct Turns Adapted To (*treated as Incorrect*)	100%	0%	36%
Problem2			
Ave. # Turns	16.50	16.80	16.25
Ave. # Correct Turns	14.60	14.35	14.10
Ave. # Uncertain Turns	3.30	3.15	3.65

Table 2: Differences in Student Answer Attributes across Problem and Condition

and posttest scores as our learning metric. For example, in a two-way ANOVA with condition by repeated test measures design, there was a significant main effect for test phase ($F(1,57) = 34.88$, $p = 0.000$, $MSe = 0.032$), indicating that students in all conditions learned a significant amount during tutoring. However, there was no significant interaction effect between condition and test phase, indicating that how much students learned was not dependent on condition. Based on our results, we hypothesize that tutoring only one physics problem was not enough to enable our uncertainty adaptation to yield significant learning differences as measured by our pretest and posttest. We are now running a larger version of this experiment where students are tutored in five physics problems.

Our second analysis of system performance uses the isomorphic second physics problem as another dialogue-based "test". In particular, we analyze how the uncertainty adaptation in the first problem impacted the quality and quantity of student answers in the isomorphic second problem. This analysis, which is ongoing, compares conditions on a variety of dialogue-based performance metrics extracted from the first and/or second problems. As one example, we have already investigated student answers in the second problem to only those tutor questions that were answered as correct+uncertain in the first problem, were adapted to with the uncertainty adaptation, and then were repeated in the test problem. In other words, we specifically investigated how students performed on repeated questions whose original answers were correct+uncertain and adapted to. This analysis thus compared only the experimental and second control conditions. A one-way ANOVA showed that in the experimental condition, a significantly higher proportion of these correct+uncertain answers became correct+nonuncertain in the second problem, as compared to the second control condition ($F(1,33) = 4.343$, $p=.045$). This result suggests that consistently adapting to uncertainty in correct student answers can decrease uncertainty in those answers, as compared to when the adaptation is given randomly.[6]

Our analysis of system performance will also include numerous other dialogue-based metrics, including differences in correctness, uncertainty, and turn length at different discourse structure depths; this discourse structure information is automatically available in our dialogues (Forbes-Riley et al., 2008b). We compare these and other metrics across conditions in (Forbes-Riley et al., 2008a).

A second use of the Uncertainty corpus is as a resource for analyzing prosody and other linguistic features of naturally occurring user affect in human-computer dialogue, particularly for use in automatic affect detection. For example, there has been significant prior research on the prosody of elicited or acted emotions (e.g. (Oudeyer, 2002; Liscombe et al., 2003)); however, these results generally transfer poorly to naturally occurring emotions (Cowie and Cornelius, 2003; Batliner et al., 2003). Thus recent research has focused on analyzing and detecting user affect in naturally occurring dialogue (e.g. (Vidrascu and Devillers, 2005; Batliner et al., 2003; Shafran et al., 2003)). The Uncertainty corpus provides an additional resource for this active research area, because it makes available a large number of features derived from the speech files, transcripts, and log files. We have already shown that useful predictive models of student affect in general, and student uncertainty specifically, can be built using similar features available in our ITSPOKE corpora (Litman and Forbes-Riley, 2006; Ai et al., 2006).

6. Summary

We presented the publicly available Uncertainty corpus, a collection of spoken tutoring dialogues between students and an adaptive Wizard-of-Oz spoken dialogue tutoring system, in which student uncertainty was manually annotated by the human Wizard. Uncertainty was also automatically adapted to in some dialogues. We overviewed the corpus collection and contents, including speech files, transcripts, manual uncertainty and correctness annotations, and log files. We discussed possible uses of this corpus by the computational linguistics community as a novel resource for studying naturally occurring user affect and affect adaptation in complex dialogue systems.

Acknowledgments

This work is supported by the National Science Foundation (award number 0631930). This work was also done as part of the Pittsburgh Science of Learning Center which is funded by the National Science Foundation award number SBE-0354420. We thank the ITSPOKE Group for help with the design, implementation, and collection of this corpus.

[6]For this analysis, 3 subjects in the experimental condition and 2 subjects in the second control condition were removed, because they never received the adaptation. In the experimental condition, these subjects had no correct+uncertain answers, while in the second control condition, these subjects had no correct answers randomly selected for the adaptation. However, a separate analysis showed that the conditions displayed no significant difference in the total number or percentage of correct answers in the second problem. Alternative approaches to excluding these subjects are discussed in (Forbes-Riley et al., 2008a).

7. References

H. Ai, D. Litman, K. Forbes-Riley, M. Rotaru, J. Tetreault, and A. Purandare. 2006. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In *Proceedings of Interspeech*, pages 797–800, Pittsburgh, PA.

J. Ang, R. Dhillon, A. Krupski, E.Shriberg, and A. Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In J. H. L. Hansen and B. Pellom, editors, *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 2037–2039, Denver, USA.

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth. 2003. How to find trouble in communication. *Speech Communication*, 40(1-2):117–143.

A. Batliner, C. Hacker, S. Steidl, E. Noth, and J. Haas. 2004. From emotion to interaction: Lessons from real human-machine dialogues. In E. Andre, L. Dybkjær, W. Minker, and P. Heisterkamp, editors, *Affective Dialogue Systems, Proceedings of a Tutorial and Research Workshop*, volume 3068 of *Lecture Notes in Artificial Intelligence*, pages 1–12, Berlin. Springer-Verlag.

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

R. Cowie and R. R. Cornelius. 2003. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32.

S. Craig, A. Graesser, J. Sullins, and B. Gholson. 2004. Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29(3):241–250.

K. Forbes-Riley and D. Litman. 2008. Analyzing dependencies between student certainness states and tutor responses in a spoken dialogue corpus. In L. Dybkjaer and W. Minker, editors, *Recent Trends in Discourse and Dialogue*, pages 275–304. Springer.

K. Forbes-Riley, D. Litman, and M. Rotaru. 2008a. Responding to student uncertainty during computer tutoring: A preliminary evaluation. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems (ITS)*, Montreal, Canada, June.

K. Forbes-Riley, M. Rotaru, and D. Litman. 2008b. The relative impact of student affect on performance models in a spoken dialogue tutoring system. *User Modeling and User-Adapted Interaction*, 18(1-2):11–43, February.

J. Liscombe, J. Venditti, and J.Hirschberg. 2003. Classifying subject ratings of emotional speech using acoustic features. In *Proceedings of Interspeech/EuroSpeech*, pages 725–728, Geneva, Switzerland.

D. Litman and K. Forbes-Riley. 2006. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48(5):559–590.

P-Y. Oudeyer. 2002. The production and recognition of emotions in speech: Features and Algorithms. *International Journal of Human Computer Studies*, 59(1-2):157–183.

H. Pon-Barry, K. Schultz, E. Owen Bratt, B. Clark, and S. Peters. 2006. Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education*, 16:171–194.

I. Shafran, M. Riley, and M. Mohri. 2003. Voice signatures. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 31–36, St. Thomas, US Virgin Islands.

K. VanLehn, S. Siler, and C. Murray. 2003. Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3):209–249.

L. Vidrascu and L. Devillers. 2005. Detection of real-life emotions in dialogs recorded in a call center. In *Proceedings of INTERSPEECH*, Lisbon, Portugal.

Marilyn Walker, Rebecca Passonneau, and Julie Boland. 2001. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 515–522, Toulouse, France.