Spoken Tutorial Dialogue and the Feeling of Another's Knowing

Diane Litman
University of Pittsburgh
Pittsburgh, PA 15260 USA
litman@cs.pitt.edu

Kate Forbes-Riley
University of Pittsburgh
Pittsburgh, PA 15260 USA
forbesk@cs.pitt.edu

Abstract

We hypothesize that monitoring the accuracy of the "feeling of another's knowing" (FOAK) is a useful predictor of tutorial dialogue system performance. We test this hypothesis in the context of a wizarded spoken dialogue tutoring system, where student learning is the primary performance metric. We first present our corpus, which has been annotated with respect to student correctness and uncertainty. We then discuss the derivation of FOAK measures from these annotations, for use in building predictive performance models. Our results show that monitoring the accuracy of FOAK is indeed predictive of student learning, both in isolation and in conjunction with other predictors.

1 Introduction

Detecting and exploiting knowledge of a speaker's uncertainty has been studied in several research communities. Spoken language researchers have identified statistically significant relationships between speaker uncertainty and linguistic properties of utterances such as prosody and lexical content (Liscombe et al., 2005; Dijkstra et al., 2006; Pon-Barry, 2008). Spoken dialogue researchers in turn are studying whether responding to user states such as uncertainty can improve system performance as measured by usability and efficiency (Tsukahara and Ward, 2001; Pon-Barry et al., 2006; Forbes-Riley and Litman, 2009a). In the psycholinguistics community, uncertainty has been studied in the context of metacognitive abilities, e.g. the ability to monitor the accuracy of one's own knowledge ("Feeling of Knowing" (FOK)), and the ability to monitor the FOK of someone else ("Feeling of Another's Knowing" (FOAK)) (Smith and Clark, 1993; Brennan and Williams, 1995).

Here we take a spoken dialogue systems perspective on FOAK, and investigate whether monitoring the accuracy of FOAK is a useful construct for predictive performance modeling. Our study uses data previously collected with a wizarded spoken dialogue tutoring system, where student learning is the primary performance metric. Section 2 reviews several relevant constructs and measures from the area of metacognition. Section 3 introduces our dialogue corpus and its user correctness and uncertainty annotations. Section 4 presents our method for measuring monitoring accuracy of FOAK from these annotations, while Section 5 shows how we use these measures to build predictive performance models. Our results show that monitoring the accuracy of FOAK is indeed a significant positive predictor of learning, both in isolation and over and above other predictors. As discussed in Section 6, increasing monitoring accuracy of FOAK is thus one avenue for also potentially increasing performance, which we plan to explore in future versions of our system.

2 Feeling of Another's Knowing

"Feeling of knowing" (FOK) refers to peoples' ability to accurately monitor their own knowledge, e.g. to know whether they have answered a question correctly. Psycholinguistics research has shown that speakers display FOK in conversation using linguistic cues such as filled pauses and prosody (Smith and Clark, 1993). Of perhaps more relevance to dialogue systems, research has also shown that *listeners* can use the same cues to monitor the FOK of someone else, i.e. *"feel-*

Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 286–289,
Queen Mary University of London, September 2009. ©2009 Association for Computational Linguistics

ing of another's knowing" (FOAK) (Brennan and Williams, 1995).

To quantify knowledge monitoring, measures of *monitoring accuracy* have been proposed. For example, consider an FOK experimental paradigm, where subjects 1) respond to a set of general knowledge questions, 2) take a FOK survey, judging whether or not[1] they think they would recognize the answer to each question in a multiple choice test, and 3) take such a recognition test. As shown in Figure 1, such data can be summarized in an array where each cell represents a mutually exclusive option: the row labels represent the possible FOK judgments (Y/N), while the columns represent the possible results of the multiple choice test (Y/N).

	Recognition=Y	Recognition=N
Judgment=Y	a	b
Judgment=N	c	d

$$\mathbf{Gamma} = \frac{(a)(d)-(b)(c)}{(a)(d)+(b)(c)} \qquad \mathbf{HC} = \frac{(a+d)-(b+c)}{(a+d)+(b+c)}$$

Figure 1: Measuring Monitoring Accuracy.

Given such an array, the relationship between the correctness and the judgment of FOK for answers can be measured using the standard formulas in Figure 1: **Gamma** and the Harmann coefficient **(HC)** measure relative and absolute knowledge monitoring accuracy, respectively. We use both metrics in the performance modeling experiments described in Section 5, as both measures have been well studied and there is a lack of consensus regarding their relative benefits (e.g. (Nietfeld et al., 2006)). Note that both range in value from -1 (no monitoring accuracy) to 1 (perfect monitoring accuracy).

3 Data

This research uses a corpus of 405 dialogues between 81 students and ITSPOKE-WOZ. The corpus was collected in an experiment showing that performance improved when ITSPOKE-WOZ responded to the uncertainty in student answers, over and above correctness (Forbes-Riley and Litman, 2009a).

ITSPOKE-WOZ is a semi-automatic version of ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system), which is a speech-enhanced version of the Why2-Atlas qualitative physics tutor (VanLehn et al., 2002). Recent versions of ITSPOKE have also been enhanced to detect and respond to student uncertainty, by parameterizing the dialogue manager so that it can adapt contingently based on student correctness and/or uncertainty (to test various ways of responding to uncertainty). In ITSPOKE-WOZ, a few components (speech recognition, language understanding, and uncertainty detection) have been replaced by a human "wizard", to test the upper-bound performance of adapting to uncertainty in a spoken dialogue system (i.e. without the noise introduced by speech and language processing).

The experimental procedure was as follows: Students (1) read background material, (2) took a multiple-choice pretest, (3) worked through five problems with ITSPOKE-WOZ (1 per dialogue), (4) took a survey, and (5) took a posttest similar to the pretest. Each student turn in the corpus was annotated[2] in real-time by the wizard during the experiment, using binary tags representing student *Uncertainty* (uncertain (**U**), nonuncertain (**nonU**)) and student *Correctness* (correct (**C**), incorrect (**I**)). Figure 2 shows an annotated dialogue excerpt.

ITSPOKE-WOZ$_6$: Now let's find the forces exerted on the car in the VERTICAL direction during the collision. First, what vertical force is always exerted on an object near the surface of the earth?

USER$_6$: gravity **(nonU,C)**

ITSPOKE-WOZ$_7$: Fine. In addition to the downward force of gravity, one other force is exerted on the car in the VERTICAL direction during the collision. We haven't discussed this force yet, because we only dealt with freefalling objects before this problem. But you may recall this other vertical force from the reading. What is it?

USER$_7$: um ... the friction from the air ... or the wind **(U,I)**

ITSPOKE-WOZ$_8$: Great try, but I think you know there's a mistake in your answer. Let's work it out. We know that gravity can't be the only vertical force on the car, because ...

Figure 2: Annotated Corpus Excerpt.

4 Predictors

To derive our performance models, we use **posttest** score as the dependent measure, and include **pretest** score among the independent measures. This allows us to measure learning *gain*, not just final test score. We also include additional independent measures computed from the wizard's uncertainty and correctness annotations,

[1] Likert scale rating schemes are also possible.

[2] In prior ITSPOKE corpora, this wizard displayed interannotator agreement of 0.85 and 0.62 Kappa on correctness and uncertainty, respectively (Forbes-Riley and Litman, 2008).

as detailed below. Note that all measures are computed on a per student basis (over all 5 dialogues). Table 1 shows means and standard deviations of all measures across all 81 students.

Measure	Mean	Std. Dev.
pretest	.51	.15
posttest	.75	.14
%C	.79	.09
%U	.23	.11
Gamma	.77	.17
HC	.59	.16

Table 1: Descriptive Corpus Statistics.

The percentage of student turns annotated as correct (**%C**) and as uncertain (**%U**) normalize the raw counts of the wizard's **C** and **U** annotations. Similar measures predict learning in prior experiments by ourselves and others (e.g (Litman et al., 2009)) and thus serve as useful baselines. In our corpus, 79% of a student's turns are answered correctly on average, while 77% are answered without uncertainty.

The monitoring accuracy measures **Gamma** and **HC** were introduced in Section 2. To construct an array like that shown in Figure 1, we map the first and second rows to our uncertainty annotations **NonU** and **U**, and map the columns to our correctness annotations **C** and **I**. In (Dijkstra et al., 2006), high and low FOK/FOAK judgments are similarly associated with speaker certainty and uncertainty, respectively. Note that in our annotation scheme, **NonU** answers are either certain or neutral.

5 Results: Predicting Student Learning

Given the above measures, our first prediction experiment measures the partial Pearson's correlation between each of the independent measures and **posttest**, after first controlling for **pretest** to account for learning gain. Our goal here is examine the predictive utility of the correctness, uncertainty, and monitoring dimensions in isolation.

Table 2 shows the statistically significant results of the partial correlations. The table shows the independent measure, the corresponding Pearson's Correlation Coefficient (R), and the significance of the correlation (p). As can be seen, both monitoring measures are positively correlated with learning, with **HC** providing better predictive utility than **Gamma**. However, **%C** is even more predictive of learning than either monitoring measure. Interestingly, the uncertainty measure **%U** in and of itself does not show predictive utility in this data.

Measure	R	p
%C	.52	.00
Gamma	.36	.00
HC	.42	.00

Table 2: Partial Correlations with Posttest ($p < .05$).

Our second prediction experiment uses PARADISE to build a learning model that can potentially include multiple independent measures. As in prior PARADISE applications (e.g. (Möller, 2005)), we train the models using stepwise multiple linear regression, which automatically determines the measures to include in the model. Our goal here is to explore whether monitoring accuracy provides any added value to our correctness and uncertainty measures.

When all measures are made available for predicting learning, we see that monitoring accuracy as measured by **HC** does add value over and above correctness: the stepwise procedure includes **HC** in the model, as it significantly accounts for more variance than just including **%C** and **pretest**. In particular, the application of PARADISE shows that the following performance function provides the best significant training fit to our data ($R^2 = .71, p < .01$):

$$\textbf{postest} = .44 * \textbf{\%C} + .21 * \textbf{pretest} + .20 * \textbf{HC}$$

The equation shows each selected measure and its (standardized) weight; larger weights indicate parameters with greater relative predictive power in accounting for **posttest** variance. **%C** is significant at $p < .01$, while **pretest** and **HC** are each significant at $p < .05$, with the coefficients all positive. Like the correlations, our regression demonstrates the predictive utility of the accuracy and monitoring measures, but not the uncertainty measure. The model further shows that while correctly answering the system's questions (**%C**) is predictive of learning, also including FOAK monitoring accuracy (**HC**) significantly increases the model's predictive power.

6 Conclusion and Future Directions

This paper explores whether knowledge monitoring accuracy is a useful construct for understanding dialogue system performance. In particular, we demonstrate the utility of combining previously studied correctness and uncertainty annotations, using a measure of FOAK monitoring accuracy. Our results show that while the correctness of a user's response predicts learning, the uncertainty with which a user conveys a response does not. In contrast, the ability to monitor FOAK accuracy predicts learning, in isolation and over and above correctness. We believe that monitoring accuracy will be a relevant construct for other dialogue applications involving knowledge asymmetry, such as problem solving, instruction giving, and trouble shooting (e.g. (Janarthanam and Lemon, 2008)).

In future work we plan to use our results to inform a modification of our system aimed at improving inferred user knowledge monitoring abilities; we will better measure such improvements by incorporating FOK ratings into our testing. In addition, we recently found interactions between learning and both user domain expertise and gender (Forbes-Riley and Litman, 2009b); we will investigate whether similar interactions extend to knowledge monitoring metrics. Since our corpus contains dialogues with both uncertainty-adaptive and non-adaptive versions of ITSPOKE-WOZ, we also plan to examine whether differing dialogue strategies influence the learned predictive models. Finally, we plan to replicate our analyses in a dialogue corpus we recently collected using a fully automated version of our system.

Acknowledgements

NSF #0631930 supports this work. We thank H. Ai, P. Jordan, and the reviewers for helpful comments.

References

S. E. Brennan and M. Williams. 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*.

C. Dijkstra, E. Krahmer, and M. Swerts. 2006. Manipulating uncertainty: The contribution of different audiovisual prosodic cues to the perception of confidence. In *Proc. Speech Prosody*.

K. Forbes-Riley and D. J. 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*. Springer.

K. Forbes-Riley and D. Litman. 2009a. Adapting to student uncertainty improves tutoring dialogues. In *Proc. Intl. Conf. on Artificial Intelligence in Education*.

K. Forbes-Riley and D. Litman. 2009b. A user modeling-based performance analysis of a wizarded uncertainty-adaptive dialogue system corpus. In *Proc. Interspeech*, Brighton, UK, September.

S. Janarthanam and O. Lemon. 2008. User simulations for online adaptation and knowledge-alignment in troubleshooting dialogue systems. In *Proc. SEMdial*.

J. Liscombe, J. Venditti, and J. Hirschberg. 2005. Detecting certainness in spoken tutorial dialogues. In *Proc. Interspeech*.

D. Litman, J. Moore, M. Dzikovska, and E. Farrow. 2009. Using natural language processing to analyze tutorial dialogue corpora across domains and modalities. In *Proc. Intl. Conf. on Artificial Intelligence in Education*.

S. Möller. 2005. Parameters for quantifying the interaction with spoken dialogue telephone services. In *Proc. SIGdial Workshop on Discourse and Dialogue*.

J. L. Nietfeld, C. K. Enders, and G. Schraw. 2006. A Monte Carlo comparison of measures of relative and absolute monitoring accuracy. *Educational and Psychological Measurement*.

H. Pon-Barry, K. Schultz, E. O. Bratt, B. Clark, and S. Peters. 2006. Responding to student uncertainty in spoken tutorial dialogue systems. *Intl. Journal of Artificial Intelligence in Education*.

H. Pon-Barry. 2008. Prosodic manifestations of confidence and uncertainty in spoken language. In *Proc. Interspeech*.

V. L. Smith and H. H. Clark. 1993. On the course of answering questions. *Journal of Memory and Language*.

W. Tsukahara and N. Ward. 2001. Responding to subtle, fleeting changes in the user's internal state. In *Proc. SIG-CHI on Human factors in computing systems*.

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