Exploring Affect-Context Dependencies for Adaptive System Development

Kate Forbes-Riley
Learning R&D Ctr.
Univ. Pittsburgh
Pittsburgh, PA 15260
forbesk@pitt.edu

Mihai Rotaru
Computer Science Dpt.
Univ. Pittsburgh
Pittsburgh, PA 15260
mrotaru@cs.pitt.edu

Diane J. Litman
Learning R&D Ctr.
Computer Science Dpt.
Univ. Pittsburgh
Pittsburgh, PA 15260
litman@cs.pitt.edu

Joel Tetreault
Learning R&D Ctr.
Univ. Pittsburgh
Pittsburgh, PA 15260
tetreaul@pitt.edu

Abstract

We use χ^2 to investigate the context dependency of student affect in our computer tutoring dialogues, targeting uncertainty in student answers in 3 automatically monitorable contexts. Our results show significant dependencies between uncertain answers and specific contexts. Identification and analysis of these dependencies is our first step in developing an adaptive version of our dialogue system.

1 Introduction

Detecting and adapting to user affect is being explored by many researchers to improve dialogue system quality. Detection has received much attention (e.g., (Litman and Forbes-Riley, 2004; Lee and Narayanan, 2005)), but less work has been done on adaptation, due to the difficulty of developing responses and applying them at the right time. Most work on adaptation takes a context-independent approach: use the same type of response after all instances of an affective state. For example, Liu and Picard (2005)'s health assessment system responds with empathy to all instances of user stress.

Research suggests, however, that it may be more effective to take a *context-dependent* approach: develop multiple responses for each affective state, whose use depends on the state's context. E.g., in the tutoring domain, Pon-Barry et al. (2006) show that *human* tutors use multiple responses to uncertain student answers, depending on the answer's correctness and prior context. In the information-seeking domain, it is commonly believed that while an apology is a good default response to user frustration (as in (Klein et al., 2002)), one context requires a different response: after several frustrated user turns, the call should be forwarded to a human operator.

A context-dependent approach to affect adaptation must address 2 issues: in what contexts to adapt, and what responses to use there. This paper addresses the first issue and targets student uncertainty in our computer tutoring dialogues. Although our dialogues have a Question-Answer format, our system contains 275 tutor questions. Treating each question as a context is too labor-intensive for adaptation development and creates a data sparsity issue. Instead we treat automatically monitorable question properties as contexts. Here we examine 3 contexts: the dialogue act interpretation, and the discourse structure depth and transition, of the prior tutor question. We use χ^2 to investigate the context dependency of uncertain student answers (correct and incorrect). Our results show that some contexts are significantly associated with uncertain answers. Our next step will be to use these significant dependencies to develop system responses to uncertain answers in these contexts. These responses will be based both on our hypotheses about why uncertainty is associated with these contexts, and on analyses of human tutor responses to uncertain answers in these contexts.

2 Student Uncertainty and Prior Context

ITSPOKE is a *speech-enabled* version of a text-based tutoring system (VanLehn et al., 2002). The student first types an essay answering one of 5 qualitative physics problems. ITSPOKE parses the essay, extracts dialogue topics concerning misconceptions, then engages the student in dialogue. In this study we used 2 ITSPOKE corpora containing 4590 student turns over 185 dialogues from 37 students. Figure 1 shows an annotated dialogue excerpt.

ITSPOKE$_4$: Good. Does the packet have an acceleration? If yes please specify its direction. *[SAQ, Advance, D1]*

STUDENT$_4$: horizontally forward yes *[non-uncertain, incorrect]*

> **ITSPOKE$_5$**: Well... To find the packet's acceleration we will use Newton's 2nd law. What is the definition of this law? *[LAQ, Push, D2]*
>
> **STUDENT$_5$**: an object in motion tends to stay in motion until it's act by an outside force *[uncertain, incorrect]*
>
> **ITSPOKE$_6$**: Well... Newton's 2nd law says that the net force acting on an object equals the mass times the acceleration of the object. We can define Newton's 2nd law as F=m*a. Hence what can we determine about the acceleration of the packet given that we know the net force acting on it is not zero and points straight down? *[DAQ, Advance, D2]*
>
> **STUDENT$_6$**: I don't know *[uncertain, incorrect]*

ITSPOKE$_7$: Well... < *further discussion of Newton's 2nd law...*> So once again on the original question. Does the packet have an acceleration? If yes please specify its direction.*[SAQ, PopUp, D1]*

Figure 1: ITSPOKE Annotated Dialogue Excerpt

2.1 Uncertainty and Correctness Annotations

ITSPOKE, like most computer tutors, responds only to student correctness. ITSPOKE labels each answer as *correct* or *incorrect*[1]. If correct, ITSPOKE moves on to the next question. If incorrect, then for questions on simple topics, ITSPOKE gives the correct answer and moves on, while for questions on complex topics (ITSPOKE$_4$, Figure 1), ITSPOKE initiates a sub-dialogue with remediation questions (ITSPOKE$_5$ - ITSPOKE$_6$), before moving on.

Recent computer tutoring research has shown interest in responding to student affect[2] over correctness. Uncertainty is of particular interest: researchers hypothesize that uncertainty and incorrectness each create an opportunity to learn (VanLehn et al., 2003). They cannot be equated, however. First, an uncertain answer may be correct or incorrect (Pon-Barry et al., 2006). Second, uncertainty indicates that the student *perceives* a possible misconception in their knowledge. Thus, system responses to uncertain answers can address both the correctness and the perceived misconception.

In our ITSPOKE corpora, each student answer has been manually annotated as *uncertain* or *non-uncertain*[3]: *uncertain* is used to label answers expressing uncertainty or confusion about the material; *non-uncertain* is used to label all other answers.

2.2 Context Annotations

Here we examine 3 automatically monitorable tutor question properties as our contexts for uncertainty:

Tutor Question Acts: In prior work one annotator labeled 4 Tutor Question Acts in one ITSPOKE corpus (Litman and Forbes-Riley, 2006)[4]: *Short (SAQ), Long (LAQ), and Deep Answer Question (DAQ)* distinguish the question in terms of content and the type of answer it requires. *Repeat (RPT)* labels variants of "Can you repeat that?" after rejections. From these annotations we built a hash table associating each ITSPOKE question with a Question Act label; with this table we automatically labeled ITSPOKE questions in our second ITSPOKE corpus.

Discourse Structure Depth/Transition: In prior work we showed that the discourse structure Depth and Transition for each ITSPOKE turn can be automatically annotated (Rotaru and Litman, 2006). E.g., as shown in Figure 1, ITSPOKE$_{4,7}$ have depth 1 and ITSPOKE$_{5,6}$ have depth 2. We combine levels 3 and above (3+) due to data sparsity. 6 Transition labels represent the turn's position relative to the prior ITSPOKE turn: *NewTopLevel* labels the first question after an essay. *Advance* labels questions at the same depth as the prior question (ITSPOKE$_{4,6}$). *Push* labels the first question in a sub-dialogue (after an incorrect answer) (ITSPOKE$_5$). After a sub-dialogue, ITSPOKE asks the original question again, labeled *PopUp* (ITSPOKE$_7$), or moves on to the next question, labeled *PopUpAdv*. *SameGoal* labels both ITSPOKE RPTS (after rejections) and repeated questions after timeouts.

[1] We have also manually labeled correctness in our data; agreement between ITSPOKE and human is 0.79 Kappa (90%).

[2] We use 'affect' to cover emotions and attitudes that affect how students communicate. Although some argue 'emotion' and 'attitude' should be distinguished, some speech researchers find the narrow sense of 'emotion' too restrictive because it excludes states where emotion is present but not full-blown, including arousal and attitude (Cowie and Cornelius, 2003).

[3] A second annotator relabeled our dataset, yielding inter-annotator agreement of 0.73 Kappa (92%).

[4] Our Acts are based on related work (Graesser et al., 1995). Two annotators labeled the Acts in 8 dialogues in a parallel human tutoring corpus, with agreement of 0.75 Kappa (90%).

3 Uncertainty Context Dependencies

We use the χ^2 test to investigate the context dependency of uncertain (unc) or non-uncertain (nonunc) student answers that are correct (C) or incorrect (I). First, we compute an overall χ^2 value between each context variable and the student answer variable. For example, the Question Act variable (QACT) has 4 values: *SAQ, LAQ, DAQ, RPT*. The answer variable (SANSWER) also has 4 values: *uncC, uncI, nonuncC, nonuncI*. Table 1 (last column) shows the χ^2 value between these variables is 203.38, which greatly exceeds the critical value of 16.92 ($p \leq 0.05$, df=9), indicating a highly significant dependency. Significance increases as the χ^2 value increases.

Dependency		Obs.	Exp.	χ^2
QACT ~ SANSWER				203.38
LAQ ~ uncC	+	72	22	133.98
LAQ ~ uncI	+	43	27	11.17
LAQ ~ nonuncC	-	96	151	50.13
LAQ ~ nonuncI	=	48	60	3.10
DAQ ~ uncC	=	22	22	0.01
DAQ ~ uncI	+	37	27	4.57
DAQ ~ nonuncC	=	135	149	3.53
DAQ ~ nonuncI	=	63	59	0.35
SAQ ~ uncC	-	285	328	41.95
SAQ ~ uncI	-	377	408	17.10
SAQ ~ nonuncC	+	2368	2271	66.77
SAQ ~ nonuncI	-	875	898	5.31
RPT ~ uncC	-	7	14	4.15
RPT ~ uncI	=	22	18	1.25
RPT ~ nonuncC	-	70	98	20.18
RPT ~ nonuncI	+	70	39	33.59

Table 1: Tutor Question Act Dependencies ($p \leq .05$: critical χ^2=16.92 (df=9); critical χ^2=3.84 (df=1))

However, this does not tell us which variable values are significantly dependent. To do this, we create a binary variable from each value of the context and answer variables. E.g., the binary variable for *LAQ* has 2 values: "LAQ" and "Anything Else", and the binary variable for *uncC* has 2 values: "uncC" and "Anything Else". We then compute the χ^2 value between the binary variables. Table 1 shows this value is 133.98, which greatly exceeds the critical value of 3.84 ($p \leq 0.05$, df=1). The table also shows the observed (72) and expected (22) counts. Comparison determines the sign of the dependency: *uncC* occurs significantly more than expected (+) after LAQ. The "=" sign indicates a non-significant dependency.

Table 1 shows uncertain answers (*uncC* and *uncI*) occur significantly more than expected after LAQs. In contrast, non-uncertain answers occur significantly less (-), or aren't significantly dependent (=). Also, *uncI* occurs significantly more than expected after DAQs. We hypothesize that LAQs and DAQs are associated with more uncertainty because they are harder questions requiring definitions or deep reasoning. Not surprisingly, uncertain (and incorrect) answers occur significantly less than expected after SAQs (easier fill-in-the-blank questions). Uncertainty shows very weak dependencies on RPTs.

Table 2 shows that Depth1 is associated with more correctness and less uncertainty overall. Both types of correct answer occur significantly more than expected, but this dependency is stronger for *nonuncC*. Both incorrect answers occur significantly less than expected, but this dependency is stronger for *uncI*.

Dependency		Obs.	Exp.	χ^2
Depth# ~ SANSWER				53.85
Depth1 ~ uncC	+	250	228	5.46
Depth1 ~ uncI	-	230	283	27.55
Depth1 ~ nonuncC	+	1661	1579	24.73
Depth1 ~ nonuncI	-	575	625	12.66
Depth2 ~ uncC	-	78	101	7.80
Depth2 ~ uncI	+	156	125	11.26
Depth2 ~ nonuncC	-	664	699	5.65
Depth2 ~ nonuncI	+	304	277	4.80
Depth3+ ~ uncC	=	58	57	0.05
Depth3+ ~ uncI	+	93	70	9.76
Depth3+ ~ nonuncC	-	344	391	15.66
Depth3+ ~ nonuncI	+	177	155	4.94

Table 2: Depth Dependencies ($p \leq .05$: critical χ^2=12.59 (df=6); critical χ^2=3.84 (df=1))

At Depths 2 and 3+, correct answers occur significantly less than expected or show no significance. Incorrect answers occur significantly more than expected, and the dependencies are stronger for *uncI*. We hypothesize that deeper depths are associated with increased uncertainty and incorrectness because they correspond to deeper knowledge gaps; uncertainty here may also relate to a perceived lack of cohesion between sub-topic and larger solution.

Table 3 shows Pushes have the same dependencies as deeper depths (increased uncertainty and incorrectness); however, here the *uncI* dependency is only slightly stronger than *nonuncI*, which suggests that increased uncertainty at deeper depths is more reliably associated with remediation questions after the Push. Although uncertainty shows only weak dependencies on PopUps, after PopUpAdvs the *uncI* dependency is strong, with *uncI* occurring more than expected. We hypothesize that this dependency relates to students losing track of the original question/larger topic. Uncertainty shows only weak dependencies on Advances. After NewTopLevels, incorrect answers occur less than expected, but the dependency is stronger for *nonuncI*. After SameGoals, incorrect answers occur more than expected, but the dependency is stronger for *nonuncI*. Compared with the RPT results, the SameGoal results suggest students feel increased uncertainty after timeouts.

Dependency		Obs.	Exp.	χ^2
TRANS ∼ SANSWER				190.97
Push ∼ uncC	=	68	57	2.89
Push ∼ uncI	+	100	70	16.37
Push ∼ nonuncC	-	313	392	44.51
Push ∼ nonuncI	+	193	155	14.13
PopUp ∼ uncC	-	23	36	5.89
PopUp ∼ uncI	-	32	45	4.68
PopUp ∼ nonuncC	=	260	251	0.81
PopUp ∼ nonuncI	+	117	99	4.47
PopUpAdv ∼ uncC	=	8	13	2.50
PopUpAdv ∼ uncI	+	32	17	16.22
PopUpAdv ∼ nonuncC	-	76	93	7.72
PopUpAdv ∼ nonuncI	=	44	37	1.89
Advance ∼ uncC	=	217	205	1.70
Advance ∼ uncI	-	223	254	9.06
Advance ∼ nonuncC	+	1465	1416	8.66
Advance ∼ nonuncI	-	530	560	4.51
NewTopLevel ∼ uncC	=	53	54	0.04
NewTopLevel ∼ uncI	-	49	67	6.47
NewTopLevel ∼ nonuncC	+	463	375	57.33
NewTopLevel ∼ nonuncI	-	80	148	47.63
SameGoal ∼ uncC	=	17	21	0.70
SameGoal ∼ uncI	+	43	25	14.24
SameGoal ∼ nonuncC	-	92	152	44.25
SameGoal ∼ nonuncI	+	92	56	31.43

Table 3: Transition Dependencies ($p \leq .05$: critical χ^2=25.00 (df=15); critical χ^2=3.84 (df=1))

4 Current Directions

We analyzed dependencies between uncertain student answers and 3 automatically monitorable contexts. We plan to examine more contexts, such as a Topic Repetition variable that tracks similar questions about a topic (e.g. gravity) across dialogues.

Our next step will be to use the significant dependencies to develop system responses to uncertain answers in these contexts. These responses will be based both on our hypotheses about why uncertainty is significantly associated with these contexts, as well as on analyses of human tutor responses in these contexts, using our human tutoring corpus, which was collected with our first ITSPOKE corpus using the same experimental procedure.

We also plan to investigate context dependencies for other affective states, such as student frustration.

Acknowledgments

NSF (#0631930, #0354420 and #0328431) and ONR (N00014-04-1-0108) support this research.

References

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

A. Graesser, N. Person, and J. Magliano. 1995. Collaborative dialog patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology*, 9:495–522.

J. Klein, Y. Moon, and R. Picard. 2002. This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, 14:119–140.

C. M. Lee and S. Narayanan. 2005. Towards detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2), March.

D. Litman and K. Forbes-Riley. 2004. Predicting student emotions in computer-human tutoring dialogues. In *Proc. ACL*, pages 352–359.

D. J. Litman and K. Forbes-Riley. 2006. Correlations between dialogue acts and learning in spoken tutoring dialogues. *Natural Language Engineering*, 12(2).

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

H. Pon-Barry, K. Schultz, E. 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.

M. Rotaru and D. Litman. 2006. Exploiting discourse structure for spoken dialogue performance analysis. In *Proceedings of EMNLP*, Sydney, Australia.

K. VanLehn, P. W. Jordan, and C. P. Rosé et al. 2002. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proceedings of ITS*.

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.