Hedged Responses and Expressions of Affect in Human/Human and Human/Computer Tutorial Interactions

Khelan Bhatt (bhatkhe@iit.edu)
Martha Evens (evens@iit.edu)
Shlomo Argamon (argamon@iit.edu)
Computer Science Department, Illinois Institute of Technology
10 West 31st Street, Chicago, IL 60616

Abstract

We study how students hedge and express affect when interacting with both humans and computer systems, during keyboard-mediated natural language tutoring sessions in medicine. We found significant differences in such student behavior linked to whether the tutor was human or a computer. Students hedge and apologize often to human tutors, but very rarely to computer tutors. The type of expressions also differed—overt hostility was not encountered in human tutoring sessions, but was a major component in computer-tutored sessions. Little gender-linking of hedging behavior was found, contrary to expectations based on prior studies. A weak gender-linked effect was found for affect in human tutored sessions.

Introduction

How people interact with computers is of clear importance to the design of effective computer interfaces. The book *The Media Equation* (Reeves & Nass 1996) claims that people treat computer systems essentially the same as they treat people, though more recent work (Shechtman & Horowitz 2003; Goldstein et al., 2002) has raised serious questions about this conclusion. Differences between how people respond to human beings and how they respond to computers have been informally documented since the first experiments with natural language interfaces (Thompson, 1980). A better elucidation of the issues may improve intelligent systems design.

Specifically, understanding these issues better may aid in the development of more effective tutoring systems. In this paper, we study the differences between student reactions to our Intelligent Tutoring System (ITS), CIRCSIM-Tutor (Michael et al., 2003), and the human tutors on which it was modeled. Our goal is to characterize student hedges and expressions of affect and try to determine how our ITS could understand them and respond effectively.

We are motivated by experiments (Fox 1993) that suggest such differences for ITSs that carry out a natural language dialogue with the student. Fox carried out a "Wizard-of-Oz" experiment which showed students to be polite and friendly to human tutors when they met with them face-to-face, but decidedly rude to the same tutors when communicating with them over a slow computer link and told that a machine was tutoring them.

The current study has potentially important implications for the future development of our ITS. Investigation of how human tutors respond to student misery, frustration, and rage is the first step toward making systems more friendly and responsive. By contrast, our system's current response to student hedges and expressions of affect (as to any input it does not understand) is to tell the student what kind of input it is expecting. The result is dialogue like this:

Student: Clueless!
Tutor: Please respond with prediction table parameters.

Better understanding of how and when students express affect in tutoring sessions and the functions of such expressions in the discourse may lead to improvements in student modeling and hence tutoring effectiveness.

Background

Thompson's (1980) system was a pioneering natural-language interface designed to help U.S. Navy personnel load cargo onto ships. It thus attempted to delete all affective remarks, to avoid confusing the parser. Although the system was quite effective at its task, most of its affective input consisted of curses. By contrast, chat-oriented natural language interaction programs like ELIZA (Weizenbaum, 1966) and PARRY (Colby, 1975), can impress their users with simulated charm and intelligence, despite a lack of any deep understanding. Similarly, physicians experienced the natural language interface of Shortliffe's (1982) MYCIN and ONCOCIN programs as attractive, even though

input was restricted to one-word answers to questions.

The specific question of how to properly interpret student hedging in tutoring sessions was raised at the *NAACL Workshop on Adaptation in Dialogue Systems,* held as part of the 2001 meeting of the Association for Computational Linguistics. It was suggested that student hedges might provide useful information by reliably signaling student misconceptions. Our collaborators on the CIRCSIM-Tutor project at Rush Medical College are dubious about this suggestion, however. Ten years ago, after their first experiments with tutoring in cardiovascular physiology they resolved to stop commenting on hedges, because they felt that student hedging reflects personal communication styles more than any real confusion. Further experience has not changed their minds, although they respond with help and encouragement whenever they believe the student to be experiencing real distress (Bhatt 2004).

As well, there is an increasing recognition in the ITS community of the importance of affect. A full session at *Intelligent Tutoring Systems* 2002 was devoted to such issues (Aist et al. 2002; Kort & Reilly 2002; Vicente & Pain 2002). These papers all argue for the importance of responding to evidence of student distress. Our study is the first, to our knowledge, that explicitly studies student hedging of answers and expressions of affect by comparing human and computer tutorial sessions. Relevant in this context also is the recent general trend towards greater concern in the AI community with emotional aspects of intelligence, sparked mainly by the work of Breazeal and Brooks (Brooks et al. 1998; Breazeal 1998).

Goals and Hypotheses

We study response hedging and expressions of affect in human and machine tutoring sessions. This study incorporates both exploratory and hypothesis testing goals. The main exploratory questions that we investigated are as follows:

What kinds of hedged responses and expressions of affect do we see in human tutoring sessions?

What kinds of hedged responses and expressions of affect do we see in machine tutoring sessions?

How might the two kinds of tutoring interactions differ regarding student use of hedged responses and expressions of affect?

In addition, based on results in human/computer interaction (primarily Fox (1993) and Thompson (1980)), we formulate our main hypotheses:

H1a (Hedging Differs): *Student use of hedging differs depending on whether the tutor is a human or a computer system.*

H1b (Affect Differs): *Student use of affect differs depending on whether the tutor is a human or a computer system.*

The workshop discussion mentioned above also prompted us to investigate two subsidiary hypotheses about hedging, and how it may prove useful for student modeling:

H2a (Hedges Inform): *The presence of a hedge provides information regarding whether a student answer is right or wrong.*

H2b (Hedges Wrong)*: Hedged answers are almost always wrong and so provide near certain feedback for student modeling.*

Regarding the relevance of *H2b*, note that most computer tutoring systems cannot currently make use of 'weak' probabilistic information for student modeling, such as "hedged answers are 20% more likely to be wrong than non-hedged answers", but only more certain statements, such as "hedged answers are almost always wrong".

Gender-linked variation

Many previous studies, including Lakoff (1975) and Aries (1989), have reported that women hedge more than men, although interpretation of such claims is complex (Holmes 1984), since hedging can be a politeness or face-saving strategy, and not necessarily an expression of uncertainty. Of particular relevance are recent results on hedging in tutoring systems (Shah et al., 2002), which found that women hedge significantly more often than men when making initiatives in tutoring dialogues. If such differences are consistent, it should influence how tutoring systems interact with male and female students. We thus formulate:

H3a (Women Hedge): *Women hedge answers more often than men in tutoring interactions.*

Aries, Lakoff (1990) and Tannen (1990) all describe women as more likely to express emotion than men. Hence:

H3b (Women are Affectual): *Women use more affective expressions than men in tutoring interactions.*

Furthermore, Lakoff (1975) also describes women as apologizing more often. Thus we also consider whether:

H3c (Women Apologize): *Women apologize more often than men in tutoring interactions.*

Data Collection

Human/Human Tutoring Sessions

We collected transcripts of keyboard-to-keyboard human tutoring sessions (henceforth, *H/H sessions*) between students and their expert tutors on the subject of the baroreceptor reflex during November 1999. Sessions took place with the student and the tutor in separate rooms, communicating only via keyboard. The tutor for each session was either Joel Michael or Allen Rovick (both professors of physiology, the same tutor throughout each session), and the 25 subjects were paid volunteers, first year students at Rush Medical College enrolled in a physiology course. The data examined consists of over 51,000 words (over 12,000 lines) of student-tutor dialogue, from hour-long sessions (numbered K52-K76 in our corpus).

Human/Computer Tutoring Sessions

In November 2002, most of the first year class at Rush Medical College used CIRCSIM-Tutor (Michael et al., 2003) for one hour in a regularly scheduled laboratory session. Some students worked in pairs, some alone, so we wound up with only 66 transcripts (the *H/C sessions*), which we used as the basis for our findings about machine tutoring sessions. The system presents the same problems about the baroreceptor reflex as the human tutors and attempts to emulate their tutoring strategies. We have not yet attempted to analyze the differences between the single-user and paired sessions.

Methodology

Coding of Hedges

Hedges in the transcripts were hand-coded using a coding scheme based on the hedge types described in Shah's (2002) study of hedged initiatives. The first step was to examine transcripts of four H/H sessions (K52-K55) and to establish an initial categorization. This phase was performed collectively by Bhatt and Evens. Subsequently, the remaining twenty-two sessions were coded by each researcher independently. Each hedged instance was classed by one of the predefined types (Table 1). Inter-rater reliability was excellent, with a kappa of 0.97.

Following this initial coding and coder comparison, some hedge types were eliminated or aggregated into other types, and coding was standardized in all transcripts. Transcripts were electronically marked up using SGML tags, to facitlitate subsequent counting of hedges and hedge types for statistical analysis. The final list of hedge types, along with counts and examples of usage, is given in Table 1.

Table 1: Final list of hedge categories with definitions or examples of usage, with counts of occurrences as answers (**A**) and initiatives (**I**).

Hedge Type	A	I	Example
BELIEVE	6	0	*I believe*
EITHER_OR	2	0	*Either X orY*
EQUIVALENT	3	1	*it sounds as though*
EXPECT	12	0	*probably*
GUESS	10	1	*I guess*
KIND_OF	7	0	*Kind of*
MAYBE	4	4	*Maybe*
NOT_SURE	9	3	*I'm not sure*
Q1	61	11	Question mark after a statement
Q2	2	1	Question syntax with no "**?**"
SHOULD	1	0	*X should increase*
TAG	2	2	*It shouldn't X, should it?*
THINK	44	11	*I think*
THOUGHT	21	4	*I thought*
TRY	3	0	*I can try*

Table 2: Types of affect expressions in student responses and examples of usage, with counts of occurrences as answers (**A**) and initiatives (**I**).

Affect type	A	I	Example
AMAZEMENT	0	1	*Wow*
AMUSEMENT	0	1	*Ha ha*
APOLOGY	4	14	*Sorry*
COMPREHENSION	6	6	*I get it*
CONFUSION	1	7	*I'm a bit confused*
CONTEMPLATION	14	5	*Hmmm*
CURIOSITY	0	2	*I'm curious*
DIFFICULTY	0	2	*I'm having difficulty*
FEEDBACK	0	6	*That was helpful*
GRATITUDE	0	14	*Thank you*
GREETING	0	1	*Good morning*
PAIN	0	1	*Ouch*
REALIZATION	5	9	*Ahh*

Coding of Affect

For coding affect a similar procedure to that above was followed. Evens and Bhatt scanned the text comprising the sessions K52-K55 and searched for instances of student affect together, discussing potential instances. A set of categories was derived from these initial analyses, and the remaining sessions (K56-K76) were then coded independently by both researchers. The results were then discussed until a consensus was

reached on each instance. Table 2 lists the final categorization of the types of affect found in the data, with counts and examples. Transcripts were electronically marked up using SGML tags as above.

Parenthetically, identifying affect in student responses was quite straightforward. In fact, almost every expression of affect was explicitly signaled by the student. This is encouraging for the use of affectual cues by computer tutoring systems, since in a text-based medium it is very difficult, if not impossible, to deduce students' emotional states from implicit cues (such as sarcasm).

Results and Discussion

Hedging in Human Tutoring

Hedged answers occur on average 6.04 times per session (σ=3.77). The different kinds of hedges are given in Table 1. The two most common types by far (together accounting for more than half of all occurrences) are Q1, adding a question mark to an answer otherwise in statement form (possibly expressing a sort of "questioning intonation"), and THINK, expressing a modal likelihood assessment via grammatical metaphor.

The majority of hedged answers are correct (57.6%, N=151), and so hedging does not provide a clear-cut signal of misunderstanding on the part of the student, so the data do not support *H2b: Hedges Wrong*. However, an even larger majority of non-hedged answers are correct (80.1%, N=359). This difference is significant (one-sided p<0.001), supporting *H2a: Hedges Inform*. Indeed, wrong answers are almost twice as likely to be hedged than correct answers (42.7% versus 26.3%).

In contrast to other work, we found gender to make no significant difference in hedging answers, as women hedge answers an average of 5.46 times per session, whereas men do so 6.66 times, well within the statistical variation of our sample. Hence *H3c: Women Hedge* is not supported. No gender-linked difference was found for correctness of hedged answers either, with women and men averaging 59.1% and 56.2% correct for hedged responses, respectively.

Hedging in Machine Tutoring

Surprisingly, there was only a single hedge in all 66 H/C sessions, clearly supporting *H1a: Hedging Differs*. In this sole example the student hedges an answer with a spurious statistic "9/10" when "all", or no marker at all, would have been more correct:

S: 9/10 times the dr will dominate because the rr can't bring all the way back

Affect in Human Tutoring

Expressions of affect are fairly common in the H/H sessions; with large variations, however, between different students. Out of twenty-five sessions, twenty-two contained at least one instance of student affect, while three had none at all. The most common type is APOLOGY, with eighteen occurrences overall. Instances of affect occur 3.52 times per session (σ=2.65), with a very high level of variation between students.

Men and women express affect at similar overall rates, with average numbers of 3.66 and 3.38 occurrences per student, respectively, so *H3b: Women are Affectual* is not supported. On the other hand, although all thirteen of the sessions involving female students include at least one expression of affect, three of the male-student sessions do not. Fisher's exact test on these data gives p=0.096, so that we may (barely) reject the null hypothesis that the same fraction of men as women are likely to express affect in tutorial sessions. This supports a weaker version of *H3b*—although some men express a lot of affect, men are more likely than women to show no affect at all.

Considering just apologies (the overall most frequent expression of affect), χ^2 testing for two independent samples gives p=0.12, so the data do not permit rejection of the null hypothesis that men and women apologize at similar rates, and so we cannot support *H3c: Women Apologize*.

Affect in Machine Tutoring

There were more examples of affect than of hedging in the H/C sessions, but the 20 instances of affect found in 66 H/C sessions are still far fewer than the 88 instances found in just 25 H/H sessions. Moreover, only 12 sessions (18%) contained any affect at all, as opposed to 22 (88%) of the H/H sessions. Thus we find that our data clearly support *H1b: Affect Differs*.

Even more significant than the large difference in frequency of affect is the difference in the *kinds* of affect that students expressed when interacting with a computer system. We saw none of the kinds of affect listed in Table 2 that we found in the H/H sessions—affect-related expressions in the H/C sessions tended to be more confrontational than with a human tutor. Although some instances of affect did seem to be

genuine expressions of feeling, some seemed more designed to push and test the system. Glass (1999) reported even more hostile input to an earlier version of the system. We therefore classed such responses into 3 categories: Hostile (5 responses), Testing (4 responses), and Refusal-To-Answer (11 responses). For example, student T48 seemed to get annoyed with the system as these two "Hostile" excerpts indicate:

T: Why did you enter 'no change' for TPR?
S: you know why.
. . . .
T: Why is MAP still decreased?
S: I don't want to tell you.

T74 seems pretty annoyed too:

T: Why is MAP still decreased?
S: blalaal

However, student T60 is clearly trying to "test" the system:

T: Why did MAP change in the manner that you predicted?
S: In other words, <student's name> knows all...

So is T81, we think, but perhaps this was simple honesty:

T: Why did you enter 'no change' for TPR?
S: Nimesh said so

Conclusions

Our results clearly show strong differences in student use of hedges and expressions of affect, depending on whether they are being tutored by a human or a computer ITS. While all students hedge in sessions with human tutors, they do not hedge at all in the machine sessions (with one exception). This conclusion is also supported by experience with the Why2-ATLAS system (Rosé et al. 2002); Carolyn Rosé told us that they do not see hedging either, though they looked for it since they had also observed it frequently in human tutoring sessions (Rosé, personal communication). The progress of speech-enabled tutoring (Bratt et al. 2002) is of great interest; it is possible that a difference in communication modality can affect student hedging behavior. As well, decoding students' affect may be easier from speech, due to tonal and prosody cues (Forbes-Riley & Litman 2004).

One specific result of importance to ITS is that hedging is not a clear indication of student uncertainty or misunderstanding, as had been believed. Indeed, examination of the types of hedges most used by students leads us to believe that hedges are more connected to issues of conversational flow and politeness, rather than expression of uncertainty. This interpretation is implied by the two most common forms of hedges in our data; Q1 uses a question mark to demand a response (confirmation?) from the tutor, while THINK expresses a modal assessment via a subjective metaphor, rather than a more direct modal verb or adjunct, thus requesting that the tutor respond to the student's mental state. Further research will be needed to examine this interpretation more closely.

As opposed to hedging, students do express affect to machines, though far less often than to humans. The real difference is in the *kind* of affect expressed, though—students do not apologize to computers, nor do they thank them or give them direct feedback; they do, however, express confusion and frustration. Together with our results on hedging, this leads us to suspect that the fact that students know they are interacting with a computer changes their attitude towards the conversation, contra Reeves and Nass (1996), and they are less concerned with helping to keep the flow going than they are in 'normal' conversation (Sacks et al. 1974).

In future work, we will look at hedging and affect in more human tutoring sessions. We wonder if the fact that Michael and Rovick practice the motivational techniques described by Lepper et al. (1993) influences the fact that they receive more positive affective input. This will help us to better understand how tutor style might encourage more useful hedging and expression of affect. Currently, we are concentrating on investigating the responses made by human tutors to student expressions of distress, in order to develop rules to make CIRCSIM-Tutor more friendly and responsive.

Acknowledgments

This work was supported by the Cognitive Science Program, Office of Naval Research, under Grants No. N00014-94-1-0338 and N00014-02-1-0442 to Illinois Institute of Technology, and Grant N00014-00-1-0660 to Stanford University. The content does not reflect the position of policy of the government and no official endorsement should be inferred.

This work would have been impossible without the expert tutoring of Joel Michael and

Allen Rovick of Rush Medical College and their determination to create effective machine tutors. Thanks also to the anonymous reviewers whose expert comments helped improve this paper.

References

Aist, G., B. Kort, R. Reilly, J. Mostow, & R. Picard. (2002). Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens. *ITS 2002 Workshop on Empirical Methods for Tutorial Dialogue Systems*, San Sebastian, Spain.

Aries, E. (1989). Gender and communication. In P. Shaver and C. Hendrick, eds., *Sex and Gender, Vol. 7 of the Review of Personality and Social Psychology.* Newbury Park, CA: Sage. 149-176.

Bhatt, K.S. (2004). Classifying student hedges and affect in human tutoring sessions for the CIRCSIM-Tutor intelligent tutoring system. MS Thesis, Department of Computer Science, Illinois Institute of Technology, Chicago, IL.

Bratt, E.O., B.Z. Clark, Z. Thomsen-Gray, S. Peters, P. Treeratpituk, H. Pon-Barry, K. Schultz, D.C. Wilkins, & D. Fried. (2002). Model-based reasoning for tutorial dialogue in shipboard damage control. *Proceedings of ITS 2002,* San Sebastian, Spain, June. 63-69.

Breazeal, C. (1999). A motivational system for regulating human-robot interaction. In *Proceedings of AAAI98*, Madison, WI. 54-61.

Brooks, R., C. Breazeal, R. Irie, C. Kemp, M. Marjanovic, B. Scassellati, & M. Williamson (1998), Alternative essences of intelligence. *Proceedings of AAAI98*, Madison, WI. 961-967.

Colby, K. (1975). *Artificial paranoia.* New York, NY: Pergamon Press.

Elliott, Clark. (1998). Hunting for the holy grail With "emotionally intelligent" virtual *a*ctors, *ACM SIGART Bulletin,* 9(1) 20-28.

Forbes-Riley, K. and Litman, D. (2004). Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proceeedings of HLT-NAACL 2004,* May, Boston, MA. 201-208.

Fox, B. (1993). *The human tutorial dialogue project.* Hillsdale, NJ: Erlbaum.

Glass, M.S. (1999). *Broadening input understanding in a language-based intelligent tutoring system.* Ph.D. Thesis, Department of Computer Science, Illinois Institute of Technology, Chicago, IL.

Goldstein, M., G. Alsiö, J. Werdenhoff. (2002). The media equation does not always apply: People are not polite towards small computers. *Personal and Ubiquitous Computing. Berlin: Springer-Verlag* 6:87-96.

Holmes, J. (1984): Women's language: A functional approach. *General Linguistics* **24**(3):149-178.

Kort, B. & R. Reilly. (2002). An affective module for an intelligent tutoring system, *Intelligent Tutoring Systems* 2002.

Lakoff, R. (1975). *Language and woman's place*. New York, NY: Harper and Row.

Lakoff, R. (1990). *Talking power: The politics of language*. New York, NY: Basic Books.

Lepper, M. R., M. Woolverton, D. L. Mumme, and J-L. Gurtner. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In S. P. Lajoie and S. J. Derry (Eds.), *Computers as cognitive tools*, Hillsdale, NJ: Erlbaum, 75-105.

Michael, J.A., A.A. Rovick, A.A., M.S. Glass, Y. Zhou, & M. Evens (2003). Learning from a computer tutor with natural language capabilities. *Interactive Learning Environments, 11(3),* 233-262.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge, UK: Cambridge University Press.

Rosé, C.P., D. Bhembe, A. Roque, S. Siler, R. Shrivastava, & K. VanLehn. (2002). A hybrid natural language understanding approach for robust selection of tutoring goals. In S.A. Cerri, G. Gouardères, & F. Paraguaçu (eds.) *ITS 2002,* LNCS 2363. Berlin: Springer-Verlag. 552-561.

Sacks, H., E.A. Schegloff, & G. Jefferson. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language,* **50:**696-735.

Shechtman, N., & L.M. Horowitz (2003). Media inequality in conversation: How people behave differently when interacting with computers and people. *CHI* 5(1): 281-288.

Shah, F., M.W. Evens, J.A. Michael, & A.A. Rovick. (2002). Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions, *Discourse Processes,* 33(1) 23-52.

Shortliffe, E.H. (1982). The computer and clinical decision-making: Good advice is not enough. *IEEE Engineering in Medicine and Biology Magazine,* **1**(1), 16-18.

Tannen, D. (1990). *You just don't understand: Women and men in conversation.* New York, NY: William Morrow and Co.

Thompson, B. H. (1980). Linguistic analysis of natural language communication with computers. *Proceedings of the 8th International Conference on Computational Linguistics COLING 80)*, Tokyo, Japan, np.

Vicente, A. de, & H. Pain. (2002). Informing the detection of the students' motivational dtate: An empirical study. *Intelligent Tutoring Systems 2002*.

Weizenbaum, J. (1966). ELIZA – A computer program for the study of natural language interaction between mind and machine. *CACM* 9(1) 36-44.