Looking at the Student Input to a Natural-Language Based ITS

Chung Hee Lee[1], Martha W. Evens[1], and Michael S. Glass[2]

[1] Department of Computer Science, Illinois Institute of Technology
10 West 31st Street, Chicago, Illinois 60616, leechun@iit.edu, evens@iit.edu
[2] Department of Mathematics and Computer Science, Valparaiso University
Valparaiso, IN 46383, Michael.Glass@valpo.edu

Abstract. First year medical students at Rush Medical College used CIRCSIM-Tutor, a natural-language based ITS, in a physiology laboratory in November 2002. Analysis of the 66 hour-long machine sessions revealed many of the syntactic and spelling errors that we expected, along with a number of occasions when students did not understand the question that the tutor asked. In an attempt to encourage students to explain their causal reasoning, we added some open questions – the only student input that the system does not try to parse. This effort was at least partially effective. Almost all of the students answered some open questions seriously, and they produced much longer answers than to the ordinary tutor questions, though many stopped trying when they realized that the system was not trying to understand these answers. We also discuss some issues regarding student hedges and initiatives.

Key words: tutoring dialogue, student input, hedges, student initiatives, student affect

1. Introduction

Our original goal in building the CIRCSIM-Tutor system was to demonstrate that it is feasible to build a system that uses natural language interaction in solving problems as its main approach to tutoring and to demonstrate that this approach is effective. In fact, students show significant learning gains in pre- and post-tests; they report that they feel they learn from the system; they come back to the laboratory to use it on their own time; and they ask for copies to take home (Michael et al., 2003, Evens and Michael, to appear). In November 2002 we carried out an experiment comparing the learning gains made by students using the CIRCSIM-Tutor system with those made by students reading a carefully chosen and edited text. It showed, as we hoped, that the 40 students who used CIRCSIM-Tutor for an hour made significantly greater learning gains than the 33 who read the text. Actually 26 of the students in the control group chose to also come and use CIRCSIM-Tutor in the laboratory the next day or the day after. We wound up with 66 hour-long machine sessions. What is more, over 80% of the students completed all eight problems as compared to about 60% in the experiment in November, 1999. In this paper we describe the language input by the students in these 66 sessions and describe some of the problems in understanding and responding to this student language. Much of this language was very terse. Students tend to respond to focused questions with single words or phrases. There are abbreviations everywhere. When the students run out of official medical abbreviations, they invent their own. In Section 2 we present a fragment of a session and describe some of the recent changes made to the system. We provide some overall statistics about the number of student inputs and characterize the range of inputs to ordinary system questions in Section 3. Then in Section 4 we discuss spelling errors and spelling correction. In Section 5 we describe some of the open questions added to the system for this experiment and the student answers. In Section 6 we raise some issues involving hedges and initiatives.

2. A Snapshot of a Dialogue between Student and Machine

The system begins by asking the student to make predictions about the qualitative changes in seven important parameters of the cardiovascular system. The system marks any errors in red and launches a tutorial dialogue designed to help the student figure out what changes to make. The following is a short example:

T: What are the determinants of Stroke Volume?
S: CO, HR
T: It is true that CO=SV*HR, but what I am asking is what determines how much blood is ejected from the heart each time it beats (the sv)?
S: VENTRICULAR FILLING
T: Do you know what parameter in the prediction table determines filling?
S: CVP
T: Right, Central Venous Pressure determines filling, therefore determines Stroke Volume.
T: The other determinant is neurally controlled. What is the other determinant of Stroke Volume?
S: TPR
T: No, Total Peripheral Resistance is not a determinant. Stroke Volume is determined by Central Venous Pressure and Inotropic State. Which determinant is dominant in this case?

Before we ran the experiment we created a new version of CIRCSIM-Tutor, Version 2.9, that corrected many of the problems that appeared in the last major experiment in November, 1999. This version also gave more and better hints and asked a number of open questions, described below. The spelling correction routine, adapted from (Elmi, 1998), worked well on single words but not on phrases. We modified it to look for phrasal possibilities first.

When the system cannot understand the input, most commonly the student is lost and does not understand what the system was attempting to elicit. The system thus issues category messages. The five category messages listed below actually appeared in these sesions; the number in parentheses records the number of actual occurrences in the 66 sessions:

Please indicate increased, decreased, or unchanged. (18)
Is the mechanism of control neural or physical? (24)
Please respond with prediction table parameters. (61)
Please indicate a stage: DR, RR, or SS. (17)
Please indicate directly or inversely related. (9)

Reading the transcripts of the machine sessions from Fall, 1999, revealed that the system really short-changed the stronger students. When the student made no prediction errors the system provided no tutoring. It just proceeded to the next stage or next problem. Expert tutors, when faced with no errors to tutor, often ask open questions about the functioning of the baroreceptor reflex or ask the student to make generalizations about the problem-solving process. We had always avoided making the system ask such open questions for fear that it would not be able to parse the answers.

We decided to kill two birds with one stone: we could both provide a greater challenge to the student and collect linguistic data for extending the parser in the future by asking such open questions in the dialogue. The system, without parsing the answer, rolls out a "canned" expert answer in response. Although a canned response is a poor substitute for giving the tailored critique of the student answer, this would ensure that they see a correct answer. We thus obtained longer and richer dialogues with a large number of attested examples of lengthy student answers.

This ruse, asking questions without parsing the answers, was only partly successful. A number of students realized that the system was not parsing their answers and the result was some interesting testing behavior and some expressions of affect.

3. Overview of Student Inputs in November 2002

We obtained 66 transcripts from machine sessions on November 11 and 12, 2002, in a regularly scheduled laboratory. There were 40 students who had not been part of the control group doing the reading over the weekend. There were 33 in the control group. Most, but not all, of the students in the control group chose to come to the laboratory as well, so we wound up with 66 transcripts. Table 1 displays the number of student inputs, the number of open questions seen by this student, the number of error messages, the number of spelling errors and the number of problems completed by the student. During the laboratory session each student took a pretest, worked with CIRCSIM-Tutor, and then took a post-test. We measured learning gains from the differences between the pre-test and the post-test. These learning gains are not reported here.

We were pleased to see that there was a definite gain in terms of the number of problems completed. As shown in Table 2, 54/66 (or 81.2%) transcripts include all 8 problems. Five students did more than eight problems; that is they repeated problems, a phenomenon that we had never seen before. By comparison only 60% of the transcripts from Fall 1999 included all eight problems (21/35). The two students who did only one problem left the laboratory after about fifteen minutes. Our impression from observing students during the laboratory sessions is that the students who did fewer than eight problems were not forced to stop by our time limit but chose to stop because they felt they had learned what they could from the system.

Table 1. Statistics of Student Inputs in November, 2002

	Student Inputs	Open Qs Asked	Open Qs Answered	Category Messages	Spelling Errors	Problems Completed
Per Session	**45.2**	**8.1**	**5.9**	**2.0**	**1.6**	**7.4**
Total	**2980**	**535**	**390**	**130**	**106**	**487**

The total number of student inputs was 2980, so the average was 45.2 inputs per session. Because the focus of the dialogue is the baroreceptor reflex, the negative reflex system that controls blood pressure in the human body, many questions involve qualitative changes in the important parameters. Thus, appropriate student answers tend to involve parameter names, verbs of change and directional adverbs, adjectives that answer questions about whether a relationship is direct or inverse, and though, these are less frequent, answers to yes/no questions. The most common ways of indicating change are shown in Table 3.

Relationships between physiological parameters are typically typed by the student as “direct, dir, di, d, +, inverse, indirect, ind, inv, in, I,” and “- .” This means that we have some genuine ambiguity, since “ I, in, +, d,” and “-” may sometimes tell us about a direction of change and sometimes about a relationship.

Table 2. Numbers of Problems Completed in 2002

No. of Problems Completed	No. of Students
1	2
2	0
3	0
4	1
5	2
6	3
7	4
8	49
more than 8	5
Total	66

Table 3: How Students Indicate Change

increased, inc, in, i, I, +, up
decrease, decreased, dec, de, d, D, -, down, less
unchanged, unch, unc, 0, o, no change, same

Our ability to interpret these answers is totally dependent on remembering what question the system asked. We implemented a small ontology (Glass, 1999, 2000, 2001) for answering mechanism questions, since students seem to use almost any component or aspect of the nervous system to indicate a neural mechanism. We are now trying to expand this ontology for use in reasoning about answers to open questions (Lee et al., 2002a, 2002b). Students sometimes spell out parameter names, as in "Central Venous Pressure," more often they are abbreviated, e.g. "CVP", or replaced by a synonym such as "preload." When the system asks a question that involves several variables at once, the student sometimes types a list with "and" or commas or spaces or even an arithmetic operator as separator, so "SV HR" or "SV and HR" or "SV, HR" or "SVxHR" are all attested ways for specifying the two variables SV and HR.

4. Spelling Correction in CIRCSIM-Tutor

The students made 106 spelling errors – fewer than 2 per session. Over the last ten years their ability to type on a computer keyboard has improved markedly, making spelling correction a less overwhelming problem than it was when this project began. It is still important to any project like this, however. Students do not want to worry about making spelling corrections in the middle of solving a complex problem. The system corrected 104 out of these 106 spelling errors without making any miscorrections that we could identify, but it did miss two corrections that it should have made. It failed to correct "soconstriction" to "vasoconstriction" and "lood volume" to "blood volume."

The tutor also rejected two answers that should have been recognized as correct, because of a lack of vocabulary. The system turned down "calcium" as a mechanism, when in fact calcium ions cause the Inotropic State to increase. We have now added "calcium" to the ontology of neural mechanisms, as well as "less" and "more" as meaning decrease and increase.

These 66 sessions contain 130 category messages or almost two per session, with the prediction table message by far the most frequent. Of these 130 messages 95 were correctly interpreted by the student, 35 were not. These 35 misinterpretations were all made by a set of 12 students. Some of these students seem to be trying to correct their answer to a previous question and to be so focused on this task that they did not read the category message. An expert human tutor would almost certainly recognize this situation and accept the correction with enthusiasm; our system should do this too.

5. Open Questions and Student Answers

The list of open questions added to CIRCSIM-Tutor for this experiment (Version 2.9) is shown in Table 4. During most of the dialogue the system parsed the student input and would not let a student get away without answering a question, but the system did not even attempt to parse the student responses to open questions. The system did not ask the student one of these questions until that student entered a complete column of correct predictions, so most students did not see one of these questions until fairly late in the session. The largest number of open questions that a student could see without repeating problems was 11. The number of open questions that the system asked is shown in Table 1; as are the number of serious answers provided by the students. As the table shows, 73% of open questions elicited serious responses; most of the students took these questions seriously at first and tried to answer them. Only two students did not answer any open questions at all. There were also 125 blank responses and 20 other responses. Some students gave up answering these questions when they realized that the system was not parsing this part of their input. Others continued to answer these questions to the end. We will describe the serious answers here and defer discussion of the others until the next section.

Table 4. List of Open Questions Asked by CIRCSIM-Tutor in November, 2002.

(DR, RR, and SS identify the phases in the problem-solving process where this question is appropriate.)

DR:	Why did you predict that IS would not change?
	Can you explain why HR did not change?
	Why did you enter 'no change' for TPR?
RR:	Why did you predict that IS would increase?
	Why did you predict that HR would increase?
	Why did you predict that TPR would increase?

	Why did you predict that CVP would decrease?
	Why did MAP change in the manner that you predicted?
SS:	SV increased in DR and decreased in RR. Why did you
	predict that it would increase in SS?
	CO decreased in DR and increased in RR. Why did you
	predict that it would decrease in SS?
	Why is MAP still decreased/increased? <whichever applies>
	What does the baroreceptor reflex do?
	CVP increased in DR and decreased in RR. Why did you
	predict that it would increase in SS?

The behavior of Student M48 was typical. He gave a serious try at answering the open questions, as illustrated below. The first, second, and fourth answers are pretty good; the third is not so good.

T: What does the baroreceptor reflex do?
S: try and maintain MAP

T: Why did you enter 'no change' for TPR?
S: it is under autonomic control

T: Why did MAP change in the manner that you predicted?
S: cardiac output and hr increase3d

T: Why did you predict that CVP would decrease?
S: co increased

These answers are far from verbose; they are definitely shorter than the ideal answers written by the professor and reeled out by the system, but they certainly display more syntactic sophistication (three sentences and a verb phrase) than this student's answers to ordinary questions. What is more, they potentially reveal much more about the student's thought processes.

There were 47 answers to the first question listed in Table 4. There were 11 complete sentences ("There has not been a baroreceptor reflex yet"), 7 complete because clauses ("Because only the baroreceptor firing rate directly affects IS."), 7 noun phrases ("part of the baroreceptor reflex"), 12 responses consisting of a negative plus a noun phrase ("no reflex", "no reflex response yet"), 3 prepositional phrases ("under neural control"), and 7 other fragments ("baroreceptor not yet activated"). It is clear that we need to try to parse these inputs and determine how to respond. One option is to add more cascaded finite-state machines to the existing parser, since although these inputs are syntactically and semantically more complex than those that CIRCSIM-Tutor is parsing now, they are relatively short. We could also look at an LSA approach like that used in Auto-Tutor (Graesser et al., 1998; Person et al., 2000, 2001). Lee (2003, 2004) is working on an HPSG parser to possibly handle this problem.

6. Hedges and Student Initiatives

Students hedge to human tutors all the time. They stick in "perhaps" or "maybe" or "I think" or "I guess"; they add question marks to declarative sentences. Bhatt et al. (2004) identified 218 hedges (151 hedged answers and 67 hedged initiatives) in 25 human tutoring sessions from November 1999.

All the students hedged but the number of hedges in a session varied widely from 2 in one session to 22 in another. During an ITS session at a workshop at ACL 2001 there was a discussion of the possibility that hedges might provide useful clues to models of student knowledge or student affect. This seems unlikely since Bhatt et al. have found that, although hedged answers are more likely to be in error than answers that are not hedged, more than half of hedged answers are, in fact, correct.

These results seem to confirm the perceptions of our expert tutors. After the first eight tutoring sessions in 1989, Michael and Rovick decided to stop responding to hedges on the grounds that hedges seemed to say more about the student's preferred style of communication than about the state of the student's knowledge of the subject

matter. They have continued to respond in those cases where the student indicated some serious distress or confusion, however.

But as much as students hedge during dialogue with human tutors, we have not before observed them to hedge when conversing with the computer tutor. Although we have analyzed a number of issues involved in parsing hedges (Glass, 1999), and they are theoretically ignored during parsing of normal answers, we have no experience with them. Thus we were startled by this exchange:

T: SV increased in DR and decreased in RR. Why did you predict that it would increase in SS?
S: 9/10 times the dr will dominate because the rr can't bring all the way back

This answer is semantically correct, if syntactically incomplete, except that the quantifier/qualifier "9/10 times" is not justified by anything that the student has seen in the course. The student's answer is hedged.

We are left with some interesting questions. Is hedging really an expression of uncertainty or is an interpersonal expression of politeness or deference – politeness and deference that are due to human tutors but not to machines. Have students avoided hedging to CIRCSIM-Tutor because its language performance does not match human standards? Can we expect to see more hedges as the system's conversational performance improves? Is hedging really an unconscious expression of uncertainty or is it a conscious conversational move that expresses an interpersonal relationship?

Student initiatives are defined by Shah et al. (2002) as inputs by the student that are not intended to answer the question asked by the tutor. In the human tutoring sessions such inputs are typically explanations by the student or questions about the physiology or even challenges to whatever the tutor last said. These inputs are intended to take over the course of the dialogue, so they are truly conversational initiatives as well. In the machine sessions, in addition to many blank answers, there were are twenty inputs from students in answers to open questions that were not apparently intended as serious answers to the question; all are listed in Table 5. They seem to be expressions of affect or tests of the system by the student intended to verify that the system is not paying attention to these inputs. The open question "Why did you enter no change for TPR?" received the answers: "You know why" and "Nimesh said so." The question " Why is MAP still decreased?" was answered with "I don' t want to tell you." and "Blalaal." Another student answered an open question with the canned answer for the previous question and then answered the next question by telling the system that she knows all. Recent work on ITS emphasizes the importance of understanding and responding to student affect (Aist et al., 2002; Vicente and Pain, 2000).

Should we call these inputs initiatives or not? They are initiatives in the sense that they are not intended to carry the conversation forward in the direction that the tutor is going, but they are not intended to take over the direction of the dialogue, either – they are produced only because the student believes that they will not be understood. These inputs raise another difficult question. How can the system recognize them and how should the system respond to them when it does recognize them?

7. Conclusion

In many ways the results of the 2002 experiment were very encouraging. The system did not get caught in any of the confusions that turned up in earlier experiments. The students made impressive learning gains and expressed enthusiasm in the survey. None of the students felt impelled to curse the system. We would like to believe that this was because the system is definitely less frustrating to use, but it may just have been due to the number of observers present.

This experiment left us with a number of problems in system implementation, however. How can we improve the language produced by the system so students will venture on more initiatives? How can we parse answers to open questions and any initiatives that students produce. How can we recognize expressions of student affect and how should we respond to them? Beyond the level of our own system problems, we wonder about the role of hedges in tutoring sessions.

Acknowledgments

This work was partially supported by the Cognitive Science Program, Office of Naval Research under Grant 00014-00-1-0660 to Stanford University as well as Grants No. N00014-94-1-0338 and N00014-02-1-0442 to Illinois Institute of Technology. The content does not reflect the position or policy of the government and no official endorsement should be inferred.

Table 5. List of Nonserious Answers to Open Questions.

(Each entry contains a session number, the tutor question and the entire student response. In session T56 the student typed the response given by the system to the previous open question.)

M52 T: Why did you predict that IS would not change?	S: 0
M52 T: Why did you predict that IS would increase?	S: 1
M52 T: What does the baroreceptor reflex do?	S: l
M52 T: Can you explain why HR did not change?	S: no
M59 T: What does the baroreceptor reflex do?	S: nothing
M65 T: Can you explain why HR did not change?	S: +
M67 T: Why did you enter ' no change' for TPR?	S: BC
T48 T: Why did you enter ' no change' for TPR?	S: you know why.
T48 T: Can you explain why HR did not change?	S: yes, i can.
T48 T: Why is MAP still decreased?	S: I don' t want to tell you.
T56 T: Why did MAP change in the manner that you predicted?	S: If we think about what we mean by DR -- a period of time BEFORE any change in neural activity has occurs, then we can see that heartrate will not change because HR is determined by sympathetic and parasympathetic activity to the hear.
T60 T: Why did MAP change in the manner that you predicted?	S: In other words, <student's own name> knows all...
T60 T: Why did you predict that IS would not change?	S: it
T65 T: CO decreased in DR and increased in RR. Why did you predict hat it would decrease in SS?	S: dr
T74 T: Why is MAP still decreased?	S: blalaal
T76 T: Why did you enter ' no change' for TPR?	S: the TPR can
T79 T: Why did you predict that IS would not change?	S: whatever
T81 T: Why is MAP still decreased?	S: asdf
T81 T: What does the baroreceptor reflex do?	S: t
T81 T: Why did you enter ' no change' for TPR?	S: Nimesh said so

References

Aist, G., Kort, B., Reilly, R., Mostow, J., and Picard, R. (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.

Bhatt, K., Agamon, S.,and Evens, M. (2004). Hedged responses and expressions of affect in human/human and human/computer tutorial interactions. *COGSCI 2004,* Chicago, IL.

Elmi, M., and Evens, M. (1998). Spelling correction using context. *Proceedings of COLING 98,* Montreal, Canada. 360-364.

Evens, M. and Michael, J. (to appear). *One-on-One Tutoring by Man and Machine.* Erlbaum.

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

Glass, M.S. (2000). Processing language input in the CIRCSIM-Tutor intelligent tutoring system. In C.P. Rosé & R. Freedman (Eds.) *Proceedings of the AAAI Fall Symposium on Building Dialogue Systems for Tutorial Applications,* Menlo Park, CA: AAAI Press. 74-79.

Glass, M.S. (2001). Processing language input for an intelligent tutoring system. In J.D. Moore, C.L. Redfield, & W.L. Johnson (Eds.), *Proceedings of Artificial Intelligence in Education.*Amsterdam:IOS Press. 210-221.

Graesser, A.C., Franklin, S., Wiemer-Hastings, P. (1998). Simulating smooth tutorial dialogue with pedagogical value. *Proc. FLAIRS 98.* Sanibel, Island, FL. 163-167.

Lee, C.H., Seu, J.H., and Evens, M. (2002a). Building an ontology for CIRCSIM-Tutor. *Proc. MAICS 2002*. 161-168.

Lee, C.H., Seu, J.H., and Evens, M. (2002b). Automating the construction of case frames for CIRCSIM-Tutor. *Proc. ICAST 2002.* 59-65.

Lee, C.H., and Evens, M.W. (2003). Interleaved syntactic and semantic processing for CIRCSIM-Tutor dialogues. *Proceedings of the Midwest Artificial Intelligence and Cognitive Science Conference MAICS'03,* pp. 69-73. Cincinnati, OH.

Lee, C.H., and Evens, M. (2004). Using selectional restrictions to parse and interpret student answers in a cardiovascular tutoring system. *Proceedings of MAICS 2004*, Schaumburg, IL. 63-67.

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

Person, N.K., Graesser, A.C., Harter, D.C., Mathews, E.C. & the Tutoring Research Group, (2000). Dialog move generation and conversation management in Auto-Tutor. In C.P. Rosé & R. Freedman (Eds.) *Proceedings of the AAAI Fall Symposium on Building Dialogue Systems for Tutorial Applications,* Menlo Park, CA: AAAI Press. 87-94.

Person, N.K., Bautista, L., Graesser, A.C., Mathews, E.C., & The Tutoring Research Group. (2001). Evaluating student learning gains in two versions of Auto-Tutor. In J.D. Moore, C.L.Redfield & W.L. Johnson (Eds.), *Proceedings of Artificial Intelligence in Education.* Amsterdam: IOS Press. 246-255.

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

Vicente, A., and Pain, H. (2000). A computational model of affective educational dialogues. In C.P.Rosé & R. Freedman (Eds.) *Proceedings of the AAAI Fall Symposium on Building Dialogue Systems for Tutorial Applications.* Menlo Park, CA: AAAI Press. 113-121.