Eliciting student explanations during tutorial dialogue for the purpose of providing formative feedback

Pamela Jordan
Learning Research and Development Center
University of Pittsburgh
Pittsburgh PA, USA, 15260
pjordan@pitt.edu

Patricia Albacete
Learning Research and Development Center
University of Pittsburgh
Pittsburgh PA, USA, 15260
palbacet@pitt.edu

Michael J. Ford
School of Education
University of Pittsburgh
Pittsburgh PA, USA, 15260
mjford@pitt.edu

Sandra Katz
Learning Research and Development Center
University of Pittsburgh
Pittsburgh PA, USA, 15260
katz@pitt.edu

Michael Lipschultz
Department of Computer Science
University of Pittsburgh
Pittsburgh PA, USA, 15260
mil28@pitt.edu

ABSTRACT

In this paper we explore the question of whether additional benefits can be derived from providing formative feedback on students' explanations given the difficulties of accurately assessing them automatically. We provide a preliminary evaluation of an approach in which students assist in interpreting their own explanations and we lay out our plans for evaluating the effectiveness of a natural-language intelligent tutoring system's feedback to that interpretation effort. The preliminary evaluation suggests that students respond well to the approach. While their interpretation assistance may be similar to an automated explanation matcher, they continue to provide explanations throughout their interactions.

Keywords

student explanations, tutorial dialogue, formative feedback

1. INTRODUCTION

Numerous studies suggest that self-explanations can be more beneficial to students than explanations from others (e.g. [3]). In the context of an automated learning environment this raises the question of whether additional benefit can be derived from providing formative feedback on any explanations the student enters when the automated understanding of those explanations remains a major obstacle. Must we be satisfied with the self-explanation effect or can and should we do more?

Previous work has attempted to recognize natural language explanations and then engage in a natural language dialogue with the student to refine and improve those explanations (e.g. [11]). And more recent work has attempted to field dialogue systems that incorporate more knowledge intensive automated recognition of students' elaborations during dialogue [4]. But so far, recognizing what the student meant is still very limited. And even if we step away from attempts at actual understanding, the performance for matching to canonical sets of answers is still relatively low (e.g. [5, 12]) compared to what can be achieved with short answer responses (e.g. [13]). Perhaps even more troubling is how sensitive students are to a system's failure to understand them [4]. Although a system can recover and move forward in a coherent manner, the students notice the lack of understanding. One possibility for this sensitivity may be that the errors are often quite different from those a human makes (e.g. the system fails to recognize a response as correct but a human clearly would).

Related work, which studied the impact of decisions about dialogue tactics [2], seems to have avoided some of these issues by substituting a human interpreter (wizard) for the automated interpreter. One goal of this substitution was to reduce the confounds of misunderstandings so that the system could focus on evaluating decision policies regarding whether to elicit or tell the explanations and justifications for statements made either by the system or the student. The human interpreter was presented with a list of canonical answers and was asked to find the best match for the student's response or to select "none of the above". There were significant differences in learning based just on varying decision policies about whether to elicit or tell the same content. This result suggests that being able to request explanations and justifications and being able to reduce the confounds of errors in matching to canonical answers has potential. But is there a practical way to include a human interpreter in a classroom setting? And how sensitive are students to problems that arise if their answer is close to correct but not a good match for any of the canonical answers?

First we will introduce the Rimac[1] system and its experimental setting and our approach for eliciting and assessing students' responses to requests for explanations/justifications. Next we will describe the data we have collected and provide a preliminary evaluation of the success of our approach for eliciting explanations/justifications. Finally, we will lay out our plans for exploring if there is value added to providing feedback on students' explanations.

2. THE RIMAC SYSTEM

Rimac is a natural-language intelligent tutoring system that engages students in dialogues that address physics concepts and principles, after they have solved quantitative physics problems. Much research has been devoted to identifying features of tutorial dialogue that can explain its effectiveness (e.g., [1]), so that these features can be simulated in natural-language tutoring systems. One hypothesis is that the highly interactive nature of tutoring itself promotes learning. Several studies indicate that our understanding of interactivity needs refinement because it cannot be defined simply by the amount of interaction nor the granularity of the interaction but must also take into consideration how well the interaction is carried out (e.g., [2]).

This need for refinement suggests that we should more closely examine the linguistic mechanisms evident in tutorial dialogue. Towards this end, we first identified which of a subset of co-constructed discourse relations correlate with learning and operationalized our findings with a set of nine decision rules which we implemented in Rimac [9]. To test for causality, we created two versions of Rimac: an experimental version that deliberately executes the nine decision rules within a Knowledge Construction Dialogue (KCD) framework, and a control KCD system that does not intentionally execute these rules. One of the rules in the experimental version of Rimac that is of particular relevance for this paper is: If the student gives no explanation for an answer to a complex question then ask for one.

Rimac was built using the TuTalk tutorial dialogue toolkit [6] but has been enhanced with additional dialogue features such as reformulation of student input (e.g., [7]). The dialogues are tutor-initiative only and are primarily short answer questions in order to keep the accuracy of automatic recognition high. However, the system does request student explanations at a few key points in the dialogues. It does not attempt automatic recognition of these responses but instead always follows-up with multiple choice answers for the explanation question and a request that the student select the best match for the explanation she just provided. In effect, we are asking the student to fulfill the role of a human interpreter for the system. We give the students brief verbal instructions for what is expected of them and why as we will explain in more detail below.

Figure 1 is an excerpt of an interaction between one of the students in our evaluation set and Rimac. This excerpt shows the system requesting an explanation at line 3. The student provides her response and then the system asks the student to select the best match for her response at line 5. The system then follows-up in lines 7-9 with formative feedback that is appropriate for what the student selected as the best match for her response.

1. T: We talked about the acceleration on the way up and on the way down, but what about at the very top of the jump? What was the direction of the acceleration at the top?
2. S: zero
3. T: Why do you think there wasn‘t any acceleration?
4. S: it had a zero velocity
5. T: Thanks! So what below is the best match for what you just entered? [menu choices: ‘Since the kangaroo is not moving, there isn‘t any acceleration | The kangaroo is changing direction, so it doesn‘t have an acceleration | I guessed | None of the above‘]
6. S: Since the kangaroo is not moving, there isn‘t any acceleration
7. T: There actually was acceleration at the top of the jump, even though the kangaroo wasn‘t moving.
8. T: The kangaroo wasn‘t moving for just a very brief moment and it was like that because its velocity was switching from upward to downward. To switch from an upward velocity to a downward velocity, the kangaroo‘s velocity must reach a point where it is 0 m/s.
9. T: Since acceleration represents the change in velocity and the velocity was changing in this situation, the kangaroo‘s acceleration must be non-zero.

Figure 1: Excerpt of actual student interaction that requests student's assistance with interpretation along with the resulting formative feedback

3. THE EXPERIMENTAL DESIGN

Students in five Pittsburgh area high schools interacted with one of the two versions of Rimac during two course units (kinematics and dynamics). They used the system for one to two class periods per unit. In this paper, we examine the dialogues from the kinematics unit only.

A day or two prior to using the system, students first took a pre-test, and then completed a homework assignment in which they solved four quantitative physics problems. In a subsequent class, they used the Rimac system and finally during the next class meeting took a post-test.

Just before students began using Rimac, we introduced them to the system and read the following to them regarding requests for explanations:

> "Sometimes it will ask you to explain your response. This is regardless of whether it thinks you were right or wrong.

[1] Rimac is the name of a river whose source is in the Andes. Its name is a Quechua word meaning "talking;" hence the nickname for Rimac: "talking river." We thus considered Rimac to be well suited to a dialogue system that we intended to embed in the Andes tutoring system.

> When it asks you to explain, please be sure to type in what you were thinking that lead you to your answer. You may have to think a bit about it. If you realize that you guessed or used your intuition, that's fine; just type that.
>
> It will then follow-up with a multiple choice question and ask you to pick what is the best match for what you just wrote. It is important that you pick the best match for the explanation you just wrote and not what looks like the best explanation. Rimac needs to know what your thought process was so it can do a better job of helping you understand the physics concepts involved in solving the problem.
>
> It asks you to do this matching for explanation questions because it cannot understand explanations accurately enough. However, for all the other answers you type in it is fairly accurate."

As the student and system begin the review of an assigned homework problem, Rimac first instructs her to view a brief video that describes how to solve the homework problem and then they engage in a reflective dialogue about that problem. See [10] for a more detailed description of the pilot study and planned analyses for testing the nine interactivity decision rules. Note that in this paper we are focusing on just one of these nine rules.

4. EVALUATING THE APPROACH FOR ELICITING EXPLANATIONS

We visited the high schools at two different points during their physics courses, once when the students were ready to learn the material covered in Rimac's kinematics problems and later when they were ready for Rimac's dynamics problems. The students who used the experimental version of Rimac, in which the interactivity rules are implemented, had the opportunity to discuss three kinematics problems and three dynamics problems. In this preliminary evaluation of our elicitation approach, due to time constraints, we have examined the explanation request data for the three kinematics problems only and for just a subset of the students who used this version of the system (93 of 193 students).

From this subset of the data, we extracted and reviewed the dialogue data for 16 of the 18 system turns where students were asked to interpret their own responses. Two of the 18 turns were requests to provide equations as justifications so we excluded those for this analysis. Although there are 93 students in this data sample, not all questions are asked of all students. The questions that are asked depend on the student's responses during the dialogue. In this sampling of students there were 251 instances of requests for explanations/justifications to examine. We reviewed each response and categorized the student's initial typed response as either relevant to the question asked, an acknowledgement that she had guessed or didn't know, or irrelevant to the physics domain. We listed this analysis per problem in order to see if there is any indication that the problem, the order in which the problems were discussed or the types of questions asked might influence the students' behaviors.

1. T: Why do you think that acceleration is upward?
2. S: nevermind
3. T: Which of the following is the best match for what you just entered? [menu choices: 'Because the kangaroo changed direction | Acceleration is always in the opposite direction as velocity | I guessed | None of the above']
4. S: Acceleration is always in the opposite direction as velocity

Figure 2: Example of an actual student reaction when asked to explain an incorrect answer

	Problem 1	Problem 2	Problem 3
relevant content	88% (123)	68% (50)	72% (28)
guessed/idk	5% (9)	15% (11)	23% (9)
irrelevant content	3% (7)	16% (12)	5% (2)

Table 1: A categorization of students' responses to requests for explanations

Table 1 shows that students rarely refused to respond (3-5%) for problems 1 and 3. Refusal to answer increased noticeably for problem 2 (16%) so it does appear that the questions asked for this problem may have impacted students' willingness to respond.

While most student replies that we counted as refusals to answer were simply replies of "because", we noted in problem 2 some instances of "ok" and "never mind" as in the excerpt shown in Figure 2. Note that in this case the system is asking the student to explain her incorrect answer. After saying "nevermind" the student may be selecting what she was thinking or something that looked like a true statement. Of the 16 system turns evaluated, 8 were in response to wrong answers. However, problem 2 accounted for 6 of those requests and problem 3 for the rest. The increased refusal to answer for problem 2 may be because there are more requests to explain incorrect responses. When students self-explain and recognize they have made an error, it appears they sometimes try to figure out how to retract their answer or evade the expected follow-up. However, if the student does not realize her error, it does seem that the formative feedback could be beneficial. For example, consider the dialogue excerpt shown in Figure 3 where the student is asked to explain a wrong answer in line 6 and the formative feedback is provided in lines 8-13.

Next we evaluated how well the students performed in their role as human interpreter. We reviewed each response by following the training given to human interpreters for the experiments described in [2] which used the Cordillera system (i.e. how would I match this response to the choices provided). One of the reviewers of the students' responses (the main author) was also one of the human interpreters (wizard) for the first two versions of the Cordillera system. So it could be instructive to informally compare and contrast the interpretation tasks for the two systems and populations from the perspective of a wizard.

1. T: If the stone is thrown with a larger horizontal initial velocity, what happens to the time spent in the air (i.e. will it be the same or larger?)
2. S: larger
3. T: Hmm...So why did you think this?
4. S: more time
5. T: Thanks for explaining your reasoning.
6. T: So which of the reasons below best matches your explanation above for why you think the stone will stay longer in the air? [menu choices: ‘I used my intuition but I am not sure how to explain it in physics terms | Throwing it harder means the stone will not fall right away but will continue horizontally for a longer time before beginning to fall | Throwing it harder makes the stone spin and float on the air better | I‘ve seen this happen before | Because it is going farther, it will take longer to get there | Another reason‘]
7. S: Because it is going farther, it will take longer to get there
8. T: That doesn‘t sound right to me. What equation would you use to reason that out? [menu choices: ‘$d_x = v_x * t | d_x = v_x * t + 0.5 * a_x * t^2 | vf_x^2 = vi_x^2 + 2 * a_x * d_x$‘]
9. S: $d_x = v_x * t + 0.5 * a_x * t^2$
10. T: Almost. What is the value of a_x?
11. S: 0 m/s^2
12. T: Right, so then the equation becomes $d_x = v_x * t$.
13. T: The horizontal component of the motion is constant velocity (not accelerating), so $d_x = v_x * t$ applies. In this case, because the velocity is bigger, the displacement will be bigger, but the time will not be.

Figure 3: Excerpt of actual student interaction where system requests explanation of wrong response

Context: Problem solved for homework “A red colored stone is thrown horizontally at a velocity of 5.0 m/s from the roof of a 35.0 m building and later hits the ground below. What is the red stone’s horizontal displacement? Ignore the effects of air friction.”
Question: Why did we need to find the time first?
Choices:

1. time is the same in both directions
2. d = vt
3. we don’t have enough information to solve for displacement in the horizontal direction
4. we can find the displacement if we know how long it is moving at the given velocity
5. another reason

Figure 4: An example of where some choices offered to students for matching are related to the same underlying explanation (as in choices 1,3 and 4)

The Cordillera students were all undergraduates and their explanations were longer and required more effort to interpret and match. However, there was usually one clear candidate for the match and when matching to a correct response the criteria were that the necessary and sufficient details were present or could be easily inferred and no additional details signalled an error in thinking. The choices were authored to provide the minimum that would be needed to qualify as a complete answer. While wizards did not have to be physics experts, they did need to understand the physics concepts being discussed.

In contrast, the Rimac students were all in high school and their explanations were relatively short. We did not expect students to do well with a set of minimal match choices since we assume you need to understand the physics concepts to determine whether an answer actually matches. So instead the Rimac dialogue authors provided responses for matching that were intended to be closer to what a student might say and were based on input from teachers and responses collected during pilot testing. As a result some of the choices offered to students for matching varied only in the detail provided or how it was expressed. But these similar choices present the same formative feedback when selected. For example, in Figure 4, choice 2 is close to a good explanation but requires more detail to be complete while choices 1,3 and 4 are all related to the same underlying explanation. If the student selects 1,3 or 4 as a match then the underlying explanation is presented as an acknowledgement and may be interpreted by the student as a reformulation. If the student selects choice 2 then the system provides scaffolding that elicits the missing details.

So during our review of students’ response matching, we selected all that we considered to be potential matches and not just the best match. The rationale was that if a student selected one of a similar set of responses that had details that were missing in her response, a wizard cannot know whether the student’s self-explanation included these details and she chose not to express them or whether she thought more detail was necessary and was trying to avoid formative feedback.

After reviewing the student responses we counted the number of times we disagreed with their match choices. Again we present the results per problem. Table 2 shows that students’ performance may be similar to that of an automated explanation matcher. The larger disagreement for problem 2 could be due to students possibly trying to evade further feedback when they were asked to explain an incorrect answer or could be related to the questions or answer choices offered. If deserves a closer look in future work to see if a reason can be identified.

However, overall the students seem less perturbed by the results of their matching behaviors. They still continued to respond to the requests for explanations as shown by the

	Problem 1	Problem2	Problem 3
agree	78% (108)	59% (43)	74% (29)
disagree	22% (31)	41% (30)	25% (10)

Table 2: Reviewer agreement with students' matches of their responses

small increase in irrelevant content in Table 1, which remains low with an increase from 3 to 5% when moving from the first to last problem. The increase from problem 1 to problem 3 in "guessed/idk" could be due to fatigue, the explanations requested or more specifically asking for more explanations for incorrect answers in problems 2 and 3. Although the number of "guessed/idk" decreased from problem 2 (11) to problem 3 (9), recall that some students completed problems in two class sessions and some in one. This was because of differences in the length of classes across schools.

To give an idea of an upper bound for agreement, we do not expect 100% agreement between the reviewer and a trained human interpreter (wizard). When offline reviewers examined the selection choices made by the real-time human interpreters for the Cordillera system for just the most difficult student responses (i.e. those that fell into the "none of the above" category), the reviewer disagreed with 1% of the assignments to this category [8]. However, the lower bound that is allowable for matching when students are acting as the interpreter is still an open question. It will depend on whether formative feedback on the explanation related to their match choice is beneficial.

By the time of the workshop, we expect to have completed the above analyses for all students for the kinematics problems.

5. PLANS FOR EVALUATING THE FORMATIVE FEEDBACK GIVEN ON EXPLANATIONS

Recall that in the instructions we read to students we asked that they match the response they gave rather than picking what looks like the best response. We offer motivation to do this by pointing out that the system needs to know their thought processes so that it can provide better help for them. We are assuming that the formative feedback of a good match will be better than the "none of the above" feedback. However, this remains to be seen.

But because our experiment was not testing this specific hypothesis, we cannot answer this question directly (e.g. compare to a condition in which the formative feedback is always the "none of the above" feedback). However, we can test for correlations between various match qualities (i.e. trained reviewer agreed or disagreed with student) and learning of the concepts addressed by the requested explanation. This would suggest how important it is for students to receive more adapted formative feedback. In addition, we can test for gains on concepts covered in an explanation when the student's explanation is incorrect and relative to the quality of the match the student provided. This could suggest whether the feedback that followed was beneficial.

This preliminary analysis of the effects of formative feedback is forthcoming. We are currently scoring the pre and post-tests, which (when completed) will allow us to measure learning of particular concepts.

6. ACKNOWLEDGEMENTS

This research is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A10063 to the University of Pittsburgh. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education.

7. REFERENCES

[1] B. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13:4–16, 1984.

[2] M. Chi, K. VanLehn, D. Litman, and P. Jordan. An evaluation of pedagogical tutorial tactics for a natural language tutoring system: A reinforcement learning approach. *International Journal of Artificial Intelligence in Education*, 21:83–113, 2011.

[3] M. T. Chi, N. De Leeuw, M.-H. Chiu, and C. LaVancher. Eliciting self-explanations improves understanding. *Cognitive science*, 18(3):439–477, 1994.

[4] M. O. Dzikovska, P. Bell, A. Isard, and J. D. Moore. Evaluating language understanding accuracy with respect to objective outcomes in a dialogue system. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 471–481. Association for Computational Linguistics, 2012.

[5] A. C. Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, D. Harter, N. Person, and the TRG. Using latent semantic analysis to evaluate the contributions of students in autotutor. *Interactive Learning Environments*, 8:129–148, 2000.

[6] P. Jordan, B. Hall, M. Ringenberg, Y. Cui, and C. Rosé. Tools for authoring a dialogue agent that participates in learning studies. In *Proceeding of Artificial Intelligence in Education Conference*, pages 43–50, 2007.

[7] P. Jordan, S. Katz, P. Albacete, M. Ford, and C. Wilson. Reformulating student contributions in tutorial dialogue. In *Proceedings of 7th International Natural Language Generation Conference*, pages 95–99, 2012.

[8] P. Jordan, D. Litman, M. Lipschultz, and J. Drummond. Evidence of misunderstandings in tutorial dialogue and their impact on learning. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED), Brighton, UK, July*, 2009.

[9] S. Katz and P. Albacete. A tutoring system that simulates the highly interactive nature of human tutoring. *Educational Psychology (Special Issue on Advanced Learning Technologies)*, in press.

[10] S. Katz, P. Albacete, M. Ford, P. Jordan, M. Lipschultz, D. Litman, S. Silliman, and C. Wilson. Pilot test of a natural-language tutoring system for physics that simulates the highly interactive nature of human tutoring. In K. Yacef and H. Lane, editors,

Proceedings of Artificial Intelligence in Education Conference, 2012.

[11] M. Makatchev and K. VanLehn. Analyzing completeness and correctness of utterances using an atms. In *Proceeding of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, pages 403–410, 2005.

[12] V. Rus and A. C. Graesser. Deeper natural language processing for evaluating student answers in intelligent tutoring systems. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1495. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[13] S. Siler, C. P. Rosé, T. Frost, K. Vanlehn, and P. Koehler. Evaluating knowledge construction dialogs (kcds) versus minilessons within andes2 and alone. In *Workshop W6 on Empirical Methods for Tutorial Dialogue Systems*, page 9, 2002.