

Pilot-Testing a Tutorial Dialogue System that Supports Self-Explanation

Vincent Aleven, Octav Popescu, and Kenneth Koedinger
Human Computer Interaction Institute
Carnegie Mellon University
aleven@cs.cmu.edu, octav@cmu.edu, koedinger@cs.cmu.edu

Keywords: tutorial dialogue systems, self-explanation, evaluation of instructional systems

Abstract

Previous studies have shown that self-explanation is an effective metacognitive strategy and can be supported effectively by intelligent tutoring systems. It is plausible however that students may learn even more effectively when stating explanations in their own words and when receiving tutoring focused on their explanations. We are developing the Geometry Explanation Tutor in order to test this hypothesis. This system requires that students provide general explanations of problem-solving steps in their own words. It helps them through a restricted form of dialogue to improve explanations and arrive at explanations that are mathematically precise. Based on data collected during a pilot study in which the tutor was used for two class periods in a junior high school, the techniques we have chosen to implement the dialogue system, namely a knowledge-based approach to natural language understanding and classification of student explanation, seem up to the task. There are a number of ways in which the system could be improved within the current architecture.

Introduction

The best of the current generation of intelligent tutoring systems are more effective than classroom instruction, but not as effective as the best human tutors. How can we develop systems that help students to learn with greater understanding? Recently, many researchers have embraced the notion that developing tutorial dialogue systems is key if we want to achieve a dramatically more effective “3rd generation” of tutoring systems (Graesser, et al., 2001; Evens, et al., 2001; Rosé & Freedman, 2000; Aleven, 2001).

But what pedagogical approaches should underlie the dialogues conducted by these systems? A number of cognitive science studies have shown that self-explanation is an effective metacognitive strategy (Bielaczyc, Pirolli, & Brown, 1995; Chi, 2000; Renkl, et al., 1998). That is, when students study textbook text or worked-out examples, they learn more and with greater understanding to the extent that they explain the materials to themselves. However, not all students self-explain spontaneously and even when prompted to self-explain, it is difficult for students to arrive at good explanations (Renkl, et al., 1998).

This has led a number of researchers to investigate how self-explanation can be supported effectively by intelligent tutoring systems (Alevén & Koedinger, in press; Conati & VanLehn, 2000) or other instructional software (Renkl, in press). In previous work, we showed that even simple means of supporting self-explanation within an intelligent tutoring system, such as menus, can help students learn with greater understanding, as compared to tutored problem solving without self explanation (Alevén & Koedinger, in press). But it is plausible that students learn even more effectively if they are tutored to explain in their own words.

Our long-term goals are to find out whether a tutorial dialogue system can effectively tutor students as they produce self-explanations in their own words. Further, we want to find out whether and under what circumstances tutored natural language self-explanation is worth the time it takes, that is, helps students learn with greater understanding, as compared to the simpler approaches discussed previously. In order to achieve these goals, we are developing a tutorial dialogue system that supports self-explanation in the domain of geometry, the Geometry Explanation Tutor (Alevén, Popescu, & Koedinger, 2001). This system helps students, through a restricted form of dialogue, to produce explanations that not only get at the right mathematical idea but also state the idea with sufficient precision.

In this paper we focus on the first of our long-term goals. Based on the results of a short pilot study in a school, we discuss how far along we are towards having a robust and effective tutorial dialogue system that is ready for the classroom. More broadly, we discuss whether the techniques we have chosen to implement our tutorial dialogue system are up to the task and were a good choice.

In order to provide useful feedback to students, such a dialogue system must assess whether students' explanations are correct and sufficiently precise. It must be able to detect common errors and omissions. We have opted for a knowledge-based approach to natural language understanding, with a logic-based representation of semantics, similar in spirit to those discussed in Allen, 1995. In this regard our system is different from many other tutorial dialogue systems, which rely on statistical approaches (Graesser, et al., 2001), keyword spotting (Evens et al., 2001), or bypass language understanding altogether (Heffernan & Koedinger, 2000). As well, our system is different from many dialogue systems for applications other than tutoring, where statistical approaches or hybrid approaches combining deep and shallow methods are in vogue (e.g., Wahlster, 2001). We have presented our arguments for our choice elsewhere (Popescu & Koedinger, 2000). In the current paper we focus on the how well classification by knowledge-based NLU agrees with classification by humans.

Further, we have opted to keep dialogue management as simple as possible, adopting a "classify and react" approach described below. Many other dialogue systems, tutorial and otherwise, use more sophisticated approaches (Graesser, et al., 2001; Evens, et al., 2001; Wahlster, 2001). In part, our simple approach to dialogue management is made possible by the fact that it is reasonable to assume that each student input is an attempt at explaining. We discuss to what extent this simple dialogue management is sufficient.

The Geometry Explanation Tutor

The Geometry Explanation Tutor was built on top of an existing Cognitive Tutor (Anderson, et al., 1995) for geometry problem solving, the Geometry Cognitive Tutor™. This tutor, was developed previously by our research group together with a full-year high school geometry curriculum. The combination of curriculum and tutor have been shown to be better than “traditional” geometry classroom instruction (Koedinger, et al, 2000). It is being marketed commercially¹ and is in use in about 100 schools in the United States.

The Geometry Explanation Tutor provides for guided problem-solving practice but requires that students provide general explanations of their problem-solving steps in their own words (Aleven, et al., 2001). It helps students, through a restricted form of dialogue, to improve their explanations and arrive at explanations that are mathematically precise. Two example dialogues with the system, collected during the pilot study described below, are shown in Figures 1 and 2. So far, the dialogues deal with the topics covered in one of the units that make up the tutor curriculum, the Angles unit which deals with the geometric properties of angles.

The system’s architecture has been described elsewhere (Aleven, et al, 2001) so here we provide only broad outline. An important knowledge source is the system’s hierarchy of explanation categories, which we developed based on our analysis of a corpus of student explanations. The hierarchy consists of 149 categories, which represent the most common ways in which students express, or attempt to express, geometry theorems in their own words. For each relevant geometry rule, the hierarchy contains one or more categories representing correct and complete ways of stating the rule. For example, category COMPLEMENTARY-ANGLES-SUM-90 represents all correct and complete statements of the definition of complementary angles, including “the sum of the measures of complementary angles is 90 degrees”. For each relevant geometry rule the explanation hierarchy also contains numerous categories that represent commonly-occurring incomplete or incorrect ways of stating the rule. For example, category COMPLEMENTARY-ANGLES-90 represents sentences meaning “complementary angles are 90 degrees,” which falls just short of being a complete statement of the given geometry rule.

The system has an NLU component, primarily knowledge-based, whose task it is to construct a formal representation of the semantic content of student input and classify this representation with regard to the system’s set of explanation categories. The NLU component uses a left-corner chart parser with unification-based grammar formalism to parse the input (Rosé & Lavie, 1999). It uses the Loom description logic system (MacGregor, 1991) to construct a semantic representation of student input. The system’s domain ontology and the explanation categories are represented also in Loom. The system relies on the Loom classifier to classify the semantic representation with respect to the explanation hierarchy. This generally results in a set of explanation categories,

¹ See <http://www.carnegielearning.com>.

rather than a single category, as is illustrated in Figures 1 and 2, which show on the right the explanation categories assigned by the system to each student input.

The second main architectural component is the tutor’s production rule model. This model contains production rules that capture problem-solving strategies of an ideal student, as well as a small number of production rules that model tutorial strategy (see also Heffernan & Koedinger, 2000). These latter rules implement the following simple strategy: If the student’s explanation was classified as a complete and correct statement of an applicable geometry theorem, the tutor accepts the explanation, as illustrated in step 6 of each of the example dialogues shown in Figures 1 and 2. Otherwise, if the explanation was classified under one or more categories that represent incomplete or incorrect version of an applicable geometry theorem, the tutor selects one of these categories randomly and displays the feedback message associated with the selected category, as illustrated in many of the steps in Figures 1 and 2; for each step, the explanation category on which the tutor feedback was based is shown in bold face in the rightmost column. The system also appropriately handles explanations that are merely references to geometry rules (student gave the name of a geometry rule) and explanations focusing on the wrong rule. Since the pilot study described in this paper, a number of improvements have been made to the system, as described below.

Att.	Student Explanation	Tutor Feedback	Explanation	Categories
1.	mirror image of ARO	The Geometry Tutor is not able to understand your explanation. Sorry! Please add to your explanation or type something different. Please state a GENERAL geometry rule. Do not mention the names of the angles in the problem.	None.	
2.	mirror image	The Geometry Tutor is not able to understand your explanation. Sorry! Please add to your explanation or type something different. Please state a GENERAL geometry rule. Do not mention the names of the angles in the problem.	None.	
3.	two opposite angles are congruent	Are any two angles congruent? What tells you that the angles are congruent?	CONGRUENT-ANGLES, CONGRUENT-OPPOSITE-ANGLES	
4.	two opposite angles of an isosceles triangles are congruent	Correct, except for one thing: is any pair of angles in an isosceles triangle congruent?	CONGRUENT-ANGLES-OF-ISOSCELES-TRIANGLE, CONGRUENT-OPPOSITE-ANGLES-OF-ISOSCELES-TRIANGLE	
5.	two corresponding angles of an isosceles triangles are congruent	Correct, except for one thing: is any pair of angles in an isosceles triangle congruent?	CONGRUENT-ANGLES-OF-ISOSCELES-TRIANGLE, CONGRUENT-CORRESPONDING-ANGLES	
6.	two base angles of an isosceles triangles are congruent		CONGRUENT-BASE-ANGLES-OF-ISOSCELES-TRIANGLE	

Figure 1: First example dialogue with the Geometry Explanation Tutor

Att.	Student Explanation	Tutor Feedback	Explanation	Categories
1.	whole and parts postulate	The Geometry Tutor is not able to understand your explanation. Sorry! Please add to your explanation or type something different. Please state a GENERAL geometry rule. Do not mention the names of the angles in the problem.	None.	
2.	two complementary angles equal 90 degrees	You are going in the right direction. Exactly what kinds of angles are 90 degrees?	E Q U A L ,	ANGLES-90, COMPLEMENTARY-ANGLES-90
3.	two adjacent complementary angles equal 90 degrees	Correct, except for one thing: what exactly is 90?	COMPLEMENTARY-ANGLES-90,	ADJACENT-ANGLES, EQUAL
4.	two adjacent complementary angles equal nine degrees	Nice number. Can you state a geometry rule that justifies your answer?	COMPLEMENTARY-ANGLES, EQUAL,	ADJACENT-ANGLES, QUANTITY
5.	two adjacent complementary angles equal 90 degrees	Correct, except for one thing: what exactly is 90?	COMPLEMENTARY-ANGLES-90,	EQUAL ADJACENT-ANGLES
6.	two adjacent complementary angles add up to 90 degrees		COMPLEMENTARY-ANGLES-SUM-90,	ADJACENT-ANGLES-SUM

Figure 2: Second example dialogue with the Geometry Explanation Tutor

A Pilot Study

At the end of the 2000-2001 school year, we conducted a pilot study in order to get a sense of how well the system was working. The Geometry Explanation Tutor was used briefly in a suburban junior high school in the Pittsburgh area, as part of a 9th-grade Integrated Mathematics II course. This course covered a number of topics in geometry and algebra. Approximately 30 students of ages 14 and 15 year participated in the pilot study. The students were “honors students,” which means that within their school they were among the best of their grade level in terms of academic ability and diligence. During two 40-minute class periods, the students worked in pairs on the Geometry Explanation Tutor, running the tutor on the school’s wireless PC laptops. Earlier during the semester, the students had learned about angles and their interrelationships (the area covered in the tutor curriculum) but they did not have any computer tutoring related to these topics.

Evaluation of the student tutor interactions

The logs of the student-tutor interactions were analyzed in order to evaluate how effective the student-system dialogues were. The logs contained information about 791 explanation attempts related to 185 steps in which the application of a geometry theorem or definition was explained, or 12.3 ± 4.6 steps per pair of students. Students arrived at a complete explanation on 75% of the 185 steps. About half of the incomplete steps occurred simply because the bell rang at the end of the period. For the other half, the logs indicate that the students’ work on the problem ended abnormally. Such abnormal endings were especially likely to occur with geometry theorems that require longer

statements (angle addition and angle bisection) and seem to have been caused by long system response times.

First, we looked at the number of attempts and the time that it took students to complete their explanations. Both variables provide a measure of how difficult it is to explain problem-solving steps and how effective the system is in helping students to explain. In advance, we did not have a clear expectation what appropriate values would be or should be for these variables. However, if practice with the system helps students learn to explain, one should see a decline in the number of attempts and the time needed to explain any given theorem as students gain experience.

Overall, it took students 3.6 ± 4.5 attempts to get an explanation right. This average breaks down as follows: On the first opportunity to explain any given theorem, students needed 4.7 ± 5.6 attempts, whereas on later opportunities, they needed 2.5 ± 2.5 attempts. This decrease over time in the number of attempts to explain was observed for most of the theorems, as is illustrated in Figure 3, which provides a more detailed breakdown of the attempts needed. This figure shows also that some theorems are considerably more difficult to explain than others. Overall, the number of attempts that were needed to arrive at complete explanations seems reasonable. Further, our expectation that this number would go down as students gained experience was clearly borne out.

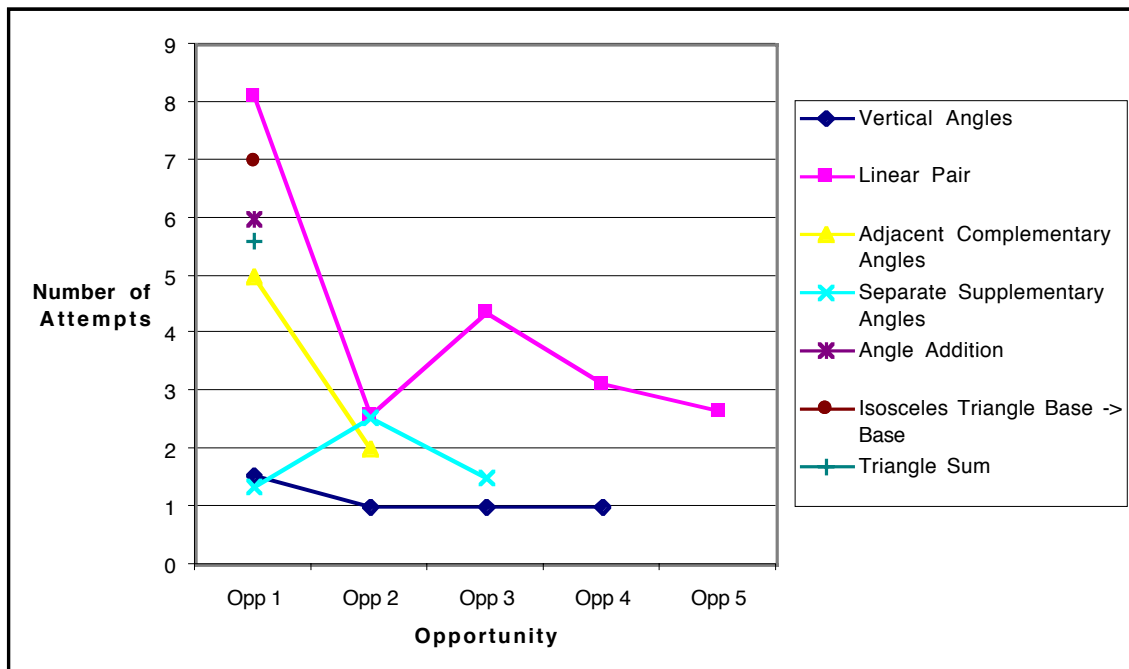


Figure 3: Average number of attempts needed to complete an explanation, per opportunity and per theorem to be explained

With regard to the second variable, time, overall, students spent 141 ± 184 seconds on each completed explanation step. This includes thinking time, typing time, and system response time. On first opportunities to explain a given theorem, students needed 196 ± 223 seconds, whereas on later opportunities, they took 84 ± 105 seconds. Thus, we see a decrease in time needed on later opportunities, as we did with the number of attempts needed to explain. The more detailed breakdown of the explanation time shown in Figure 4 illustrates that this decline is seen for most of the theorems that students explained.

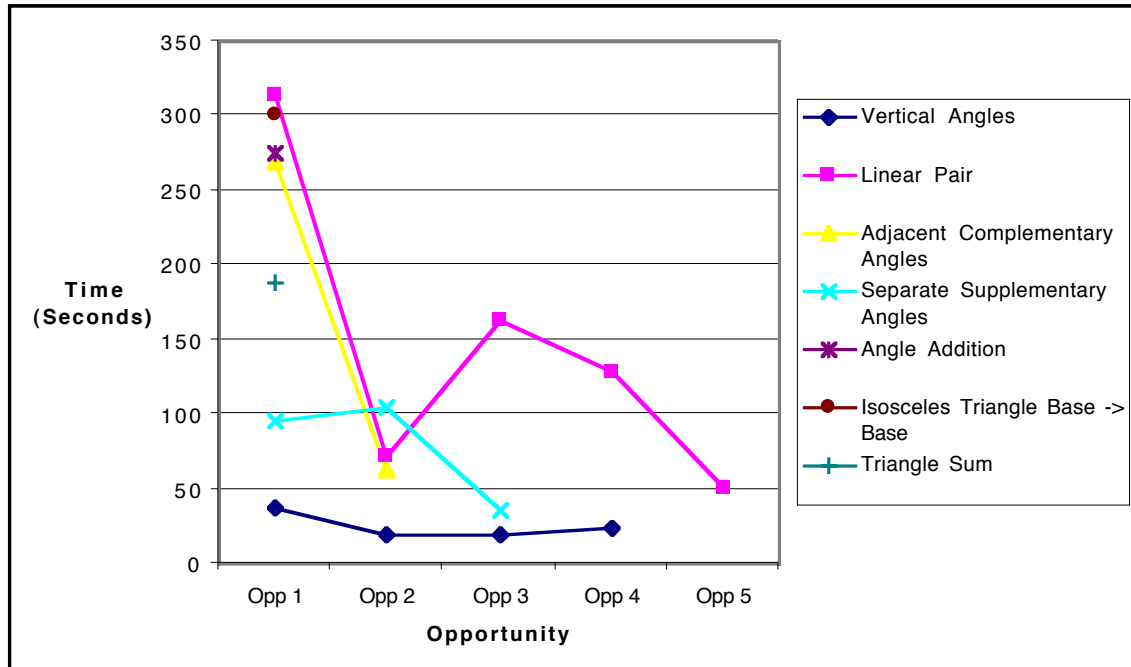


Figure 4: Average time to complete an explanation, per opportunity and per theorem to be explained.

While the average time to explain a theorem is reasonable, both on earlier and later steps. for some theorems the time was rather high, especially on first opportunities. The fact that the time needed to arrive at a complete explanation and the number of attempts declined from previous attempt provides some evidence that the students were learning to explain the geometry theorems, a positive sign.

A different way of measuring the effectiveness of the student-system dialogues is to see how often students were able to improve their explanation from one attempt to the next. In other words, to what extent does the system’s feedback guide students in incrementally improving their explanations?

We define progress as follows: an attempt at explaining a theorem is an improvement over a previous attempt if it is a correct and complete statement of an applicable geometry rule, or if it is closer to a complete statement than the previous attempt

(meaning that it classifies under a more specific category in the explanation hierarchy), or if the attempt focuses on an applicable geometry rule whereas the previous attempt focused on a geometry rule that does not apply to the current step. The actual criterion is slightly more complicated due to the fact that (a) some explanations are references to geometry rules and (b) explanations can fall under multiple categories. For example, in the dialogue shown in Figure 1, attempts 3, 4, and 6 constitute progress according to this criterion, as do steps 2, 5, and 6 in Figure 2. We define regression as the opposite of progress, see for example step 4 in Figure 2 is regression. Some explanations, constitute neither progress nor regression, even though they are classified under a different set of categories as the previous attempt. These may be explanations that are better than the previous attempt in one respect but worse in another, or explanation attempts that are different from the previous attempt but no closer to the correct explanation (see for example step 3 in Figure 2 — although not wrong, it was not necessary to add the term “adjacent”). We compute progress and regression based on the explanation categories assigned by a human rater, namely, the first author.

How often should we see incremental improvement, in order to conclude that the student-system dialogues are working well? Ideally, students would make progress on every attempt, but it seems hardly realistic to expect this level of effectiveness. (One would not expect it even from the best human tutor!). As a very rough rule of thumb, let us just say that a minimum criterion is that students make progress more often than not.

In assessing this kind of local progress, we obviously need not be concerned with the first attempt at explaining any given step (185 attempts). Further, we disregard steps where the student’s input was identical to the previous attempt. Such repetitions occurred rather frequently (namely, 218 times), but we suspect that they are mostly unintentional, due to imperfections in the system’s user interface that have been fixed meanwhile. Therefore, we focus on the remaining 388 explanation attempts in assessing local progress. Of these attempts, 44% constituted progress over previous attempts, 26% were classified under the same set of categories as the previous attempt (even though the explanation was different), 14% constituted regression, that is, explanations that were worse than the previous, and 17% of these attempts fell in none of these categories. In sum, the percentage of attempts in which students made progress (44%) was close to the 50-50 threshold presented above, but we would like it to be a bit higher.

Some changes have been made within the current dialogue management framework that are likely to improve the progress rate. First, due to its policy of selecting randomly when a student explanation classifies under multiple explanation categories, the tutor did not always base feedback on the most appropriate explanation category (i.e., did not always make good use of the information supplied by the NLU component), see for example step 2 in Figure 2. We have meanwhile changed the tutor’s selection criterion so that the tutor selects the explanation category that is closest to a complete explanation. Second, the tutor feedback can likely be improved by associating multiple levels of (increasingly specific) feedback messages with each explanation category. This enables the system to provide more specific feedback when the student’s explanation classifies under the same explanation categories as the previous attempt. This has since been implemented. Third,

some parts of the explanation hierarchy are not yet fine-grained enough to detect certain improvements to explanations. We have added 18 categories already (for a total of 167), based on the analysis of the current data set and will likely add more. Finally, it is likely to be useful if the tutor pointed out to the student whether an explanation attempt implies progress or not. The criterion for progress used here will be a good stating point. We have not yet worked on this.

While these improvements are very likely to raise the progress rate, we cannot rule out that for some of the more difficult-to-state geometry theorems, such as angle addition, which was not covered much in the current data set, we will need to model more elaborate dialogue strategies. To implement such strategies will require significant extensions to the current framework.

Evaluation of classification accuracy

We also evaluated the system's NLU component, to get an idea of how well we are doing and because the question of how well knowledge-based natural language understanding works in analyzing mathematical explanations by novices is an interesting research question in its own right. We focused on the accuracy with which the system's NLU component is able to classify student explanations with respect to its set of explanation categories.

There is inherent ambiguity in this classification task: for some explanations, it is difficult even for human raters to determine what the correct category or categories are. However, if human raters sometimes disagree, it is not reasonable to expect the system to agree all of the time with any particular human, let alone with a committee of human raters. Therefore, rather than compare the labels assigned by the system against a "correct" set of labels for each example, we ask to what extent the agreement between system and human raters approaches that between human raters. This design has been applied for example in the evaluation of Latent Semantic Analysis (a statistical technique for natural language processing) for the purpose of grading students' written essays (Foltz, Laham, & Landauer, 1999).

From the set of explanations collected during the pilot study, we removed the examples that are identical to the previous attempt. As mentioned, we strongly suspect that these repetitions were unintentional, caused by small flaws in the user interface which have been fixed. Therefore, these repetitions should not influence the evaluation of NLU performance. Three human raters labeled the remaining set of 573 examples, using the labels in the explanation hierarchy. The graders were free to assign multiple labels to each explanation. Two of the graders are authors of this paper, the third, a research assistant.

Each rater went through the data twice, in an attempt to achieve maximum accuracy. After the first round, we selected 24 "difficult" examples, explanations that all raters had labeled differently. The raters discussed these examples extensively. We then removed from the data the 24 examples and their duplicates, for a total of 31 examples, and all raters independently revised their labels. The sets of labels assigned by the system and

the human raters were then processed automatically, in order further to reduce labeling errors, inconsistencies, and irrelevant differences between label sets. These changes preserved the intention of the human raters. Also, these changes would not have affected the system's responses if they had been done on-line². The resulting agreement numbers are a more meaningful measure of the agreement among the raters.

The raters could select from a total of 167 labels, one for each of the categories in the explanation hierarchy, plus some extras for explanations that were references to geometry rules. Out of those, 91 were actually used by the human raters or the system, combined in 218 different sets of labels. To compute the inter-rater agreement, we use the κ statistic (Cohen, 1960), as is customary in the field of computational linguistics and other fields (Carletta, 1996). The κ statistic provides a measure of how much agreement there is between raters beyond the agreement expected to occur by chance alone.

We computed κ in three different ways: First we computed κ based on “set equality” – two raters were considered to agree only if they assigned the exact same set of labels to an explanation. However, this measure seems unduly harsh when there are small differences between label sets. Therefore, we also computed two versions of “weighted κ ” (Cohen, 1968), a version of the κ statistic that takes into account the degree of difference between pairs of labels (or in our case, label sets). So, second, we computed a weighted κ based on “overlap” – meaning that the degree of disagreement was computed as the ratio of the number of unshared labels versus the total number of labels. Third, we computed a weighted κ based on “weighted overlap” – to take into account a (somewhat rough) measure of semantic similarity between the individual labels. Thus we computed disagreement as a weighted sum over the labels in the two sets, where the weight of each shared label was zero and the weight of each unshared label was based on the minimum distance in the hierarchy between this label and the second set of labels. In the discussion that follows, we interpret the set equality measure as a lower bound on the agreement and focus mostly on the other two measures.

For each of the three agreement measures, we computed the average of the κ between each pair of human raters, as well as the average of the κ between the system and each human rater. The results are shown in Table 1. The average human-human κ was good. The set equality gives a lower bound of .77. According to the (more appropriate) overlap and weighted overlap measures, the human-human agreement is .81 and .88, respectively. We suspect that not all of the observed disagreement among human raters is the result of ambiguity in the data. A significant portion of it is likely to be the result of labeling errors, which are very difficult to avoid, given the large number of labels and given that multiple labels could be assigned to each explanation.

² At least if the system had used the improved criterion for selecting from among multiple labels, described in the previous section.

Table 1: Average pair-wise inter-rater agreement between human raters and average pair-wise agreement between the system and each human rater.

	κ	Actual Agreement	Chance Agreement
Set equality			
Avg Human-Human	0.77	0.77	0.033
Avg System-Human	0.60	0.61	0.030
Overlap			
Avg Human-Human	0.81	0.81	0.043
Avg System-Human	0.65	0.66	0.039
Weighted overlap			
Avg Human-Human	0.88	0.91	0.26
Avg System-Human	0.75	0.81	0.25

Overall, the system-human κ s were reasonable but lower than the corresponding human-human κ s. The κ according to the overlap measure is .65, according to the weighted overlap measure, it was .75. Thus, while the comparison of human-human κ and human-system κ indicates that the system’s classification accuracy was quite good, there seems to be some room for improvement.

In an attempt to find ways to improve the NLU component, we examined cases where there was high agreement among the human raters (i.e., at least 2 out of the 3 human raters were in full agreement, according to the set equality measure), but where the system’s classification did not agree with the majority of human raters (in terms of set equality). There were 170 such cases. A detailed examination of those cases revealed about 32 different causes for the system’s failure, ranging from difficult to very minor. The most difficult problems deal with insufficient flexibility in the face of ungrammatical language and cases where the system’s semantics model was not developed enough to deal with the complexity of the meaning of the student’s explanations. The system needs better repair capabilities to deal with ungrammatical sentences such as “the measures of the two angles in the linear pair are add up to 180 degree.” Also, the system needs a better semantic representation of coordinated structures, to handle for example the sentence “adjacent supplementary angles form a straight line and are a linear pair.” Further, a number of problems of medium difficulty need to be addressed, dealing with quantifiers, relative clauses, multi-clause sentences, and the semantics of certain predicate constructs (e.g., “isosceles triangles have two sides equal”). Finally, there are a number of small flaws with various components of the system that can easily be fixed. A number of problems have been fixed already. While we expect that it will take a considerable amount of time to address all problems, overall the evaluation results suggest that knowledge-based NLU with logic-based semantics is able to perform the detailed analysis of student explanations necessary to provide helpful feedback.

Discussion and conclusion

We are developing the Geometry Explanation Tutor in order to evaluate the hypothesis that natural language self-explanation can be tutored effectively by an intelligent tutoring system and leads to improved learning, as compared to alternative ways of supporting self-explanation. We conducted a pilot study to get an idea of the effectiveness of the current system and of the techniques chosen to implement it. This pilot study constituted the first time that the system was used in a school by a group of students from the target population and thus represents a realistic test for the Geometry Explanation Tutor. It does not yet represent a full test, given the limited time and scope; it covered about half the geometry theorems and definitions of the relevant curriculum unit.

We found evidence that the student-system dialogues are beginning to work well. The logs of student-system dialogues showed evidence that students were learning to explain geometry theorems. On the other hand, the data also revealed some room for improvement. We would like to see a higher completion rate for the dialogues. Also, the number of attempts within dialogues on which students made progress was decent but could be higher. We have described a number of measures that we took in order to improve the system's feedback, which we expect will lead to better progress rates.

Further, there was evidence that the system's knowledge-based NLU component is reaching a reasonably good level of performance in classifying student explanations. We found that the system's classification of student explanations agrees reasonably well with that of humans, no mean feat given that we are classifying with respect to a fine-grained set of categories. But the agreement between system and human raters was not as high as that between different human raters. All in all, the results are encouraging but also indicate room for improvement.

What does the evaluation say about the particular techniques we have chosen for dialogue management and natural language understanding? For some geometry theorems, such as vertical angles, linear pair, and supplementary angles, it seems quite clear that the current dialogue management framework is adequate. It is still an open question however, whether this kind of architecture is going to help students to explain such difficult-to-state theorems as angle addition. At this point, we need to leave open the possibility that we will need more elaborate dialog strategies. Knowledge-based natural language understanding with logic-based semantics seems to be able to deal with challenging input such as students' mathematical explanations.

The broader goal of our project is to get students to learn with greater understanding and do so in an efficient manner, as compared to other forms of tutored self-explanation. We are currently involved in a larger evaluation study, involving two schools, three teachers, and four classes, that focuses on this question by comparing the current system to a previous version where students explained their steps by making reference to a rule in a glossary.

References

- Aleven, V. (Ed.). (2001). Papers of the AIED-2001 Workshop on Tutorial Dialogue Systems (pp. 59-70). Available via <http://www.hcrc.ed.ac.uk/aied2001/workshops.html>.
- Aleven, V. & Koedinger, K. R. (in press). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, 26(2).
- Aleven V., Popescu, O., & Koedinger, K. R. (2001). Towards Tutorial Dialog to Support Self-Explanation: Adding Natural Language Understanding to a Cognitive Tutor. In J. D. Moore, C. L. Redfield, & W. L. Johnson (Eds.), *Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future* (pp. 246-255). Amsterdam, IOS Press.
- Allen, J. (1995). *Natural Language Understanding* (2nd Ed.). Redwood City, CA: Cummings.
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *The Journal of the Learning Sciences*, 4, 167-207.
- Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in Self-Explanation and Self-Regulation Strategies: Investigating the Effects of Knowledge Acquisition Activities on Problem Solving. *Cognition and Instruction*, 13, 221-252.
- Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249-254.
- Chi, M. T. H. (2000). Self-Explaining Expository Texts: The Dual Processes of Generating Inferences and Repairing Mental Models. In R. Glaser (Ed.), *Advances in Instructional Psychology*, (pp. 161-237). Mahwah, NJ: Erlbaum.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.
- Conati C. & VanLehn K. (2000). Toward Computer-Based Support of Meta-Cognitive Skills: a Computational Framework to Coach Self-Explanation. *International Journal of Artificial Intelligence in Education*, 11, 398-415.
- Evens, M. W., Brandle, S., Chang, R.C., Freedman, R., Glass, M., Lee, Y. H., Shim L.S., Woo, C. W., Zhang, Y., Zhou, Y., Michael, J.A. & Rovick, A. A. (2001). CIRCSIM-Tutor: An Intelligent Tutoring System Using Natural Language Dialogue. In *Twelfth Midwest AI and Cognitive Science Conference, MAICS 2001* (pp. 16-23).
- Foltz, P. W., Laham, D. & Landauer, T. K. (1999). Automated Essay Scoring: Applications to Educational Technology. In *Proceedings of EdMedia '99*.
- Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent Tutoring Systems with Conversational Dialogue. *AI Magazine*, 22(4), 39-51.
- Heffernan, N. T. & Koedinger, K. R. (2000). Intelligent Tutoring Systems are Missing the Tutor: Building a More Strategic Dialog-Based Tutor. In C. P. Rose & R. Freedman (Eds.), *Building Dialogue Systems for Tutorial Applications, Papers of the 2000 AAAI Fall Symposium* (pp. 14-19). Menlo Park, CA: AAAI Press.
- Koedinger, K. R., Corbett, A. T., Ritter, S., & Shapiro, L. (2000). *Carnegie Learning's Cognitive Tutor™: Summary Research Results*. White paper. Available from Carnegie Learning Inc., 1200 Penn Avenue, Suite 150, Pittsburgh, PA 15222, E-mail: info@carnegielearning.com, Web: <http://www.carnegielearning.com>.

- MacGregor, R. (1991). The Evolving Technology of Classification-Based Knowledge Representation Systems. In J. Sowa (ed.), *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. San Mateo, CA: Morgan Kaufmann.
- Popescu, O., & Koedinger, K. R. (2000). Towards Understanding Geometry Explanations In *Proceedings of the AAAI 2000 Fall Symposium, Building Dialog Systems for Tutorial Applications* (pp.80-86). Menlo Park, CA: AAAI Press.
- Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from Worked-Out Examples: the Effects of Example Variability and Elicited Self-Explanations. *Contemporary Educational Psychology*, 23, 90-108.
- Renkl, A. (in press). Learning from Worked-Out Examples: Instructional Explanations Supplement Self-Explanations. *Learning and Instruction*.
- Rosé, C. P. & R. Freedman, (Eds.). (2000). *Building Dialogue Systems for Tutorial Applications. Papers from the 2000 AAAI Fall Symposium*. Menlo Park, CA: AAAI Press.
- Rosé, C. P. & Lavie, A. (1999). LCFlex: An Efficient Robust Left-Corner Parser. User's Guide, Carnegie Mellon University.
- Wahlster, W. (2001). Robust Translation of Spontaneous Speech: A Multi-Engine Approach. Invited Paper, *IJCAI-01, Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* (pp. 1484-1493). San Francisco: Morgan Kaufmann.