What Makes Human Explanations Effective?

Johanna D. Moore*
Department of Computer Science, and
Learning Research and Development Center
University of Pittsburgh
Pittsburgh, PA 15260
jmoore@cs.pitt.edu

Abstract

If computer-based instructional systems are to reap the benefits of natural language interaction, they must be endowed with the properties that make human natural language interaction so effective. To identify these properties, we replaced the natural language component of an existing Intelligent Tutoring System (ITS) with a human tutor, and gathered protocols of students interacting with the human tutor. We then compared the human tutor's responses to those that would have been produced by the ITS. In this paper, I describe two critical features that distinguish human tutorial explanations from those of their computational counterparts.

Introduction

There is growing interest in teaching real world problem-solving tasks using computer-based intelligent apprenticeship environments in which students learn by doing (Gott, 1989). Such skills typically involve complex chains of hidden reasoning and one goal of an apprenticeship environment is to help externalize the cognitive processes that usually take place only mentally. Collins and Brown (1988) argue that reflection on the differences between novice and expert performance provides one means of externalizing complex cognitive processes. Moreover, psychological experimentation, e.g., (Owen and Sweller, 1985; Sweller, 1988) indicates that learning from task situations requires significant cognitive effort, and therefore some argue that much of the instruction should actually take place in post-problem *reflective follow-up* (RFU) sessions in which students review their own actions and compare them to expert behavior (Lesgold, in press). Collins and Brown (1988) further propose that computers can be a powerful tool for learning through reflection because they make it possible to represent and record the processes by which a novice or expert carries out a complex task. They argue that such a process trace, "properly abstracted and structured", can help students improve their performance on complex cognitive tasks by allowing them to systematically examine and compare their performance to that of more expert performers.

Although many have argued that reflective interactions can be an important part of the learning process, there has been no systematic attempt to develop a model of the type of dialogue that will facilitate learning through reflection. Experience with the SHERLOCK system (Lesgold *et al.*, 1992), an intelligent apprenticeship environment that trains avionics technicians to troubleshoot complex electronic devices, has shown that building a system to participate in reflective dialogues in a complex domain poses a difficult challenge. A rudimentary RFU facility has been implemented in SHERLOCK. Using this facility, students replay their solution one step at a time, and can ask the system to comment on their actions, justify its conclusions about the status of components, or explain what step an expert would have performed. SHERLOCK produces responses to these queries by filling in and printing templates selected on the basis of the question type and the particulars of the student's action and the problem situation. Due to the complexity of the domain, there is frequently a large amount of information that is potentially relevant to the student's question. Experience with the system has shown that explanations often become long and difficult to understand. This is not surprising since the current explanation facility simply patches together all of the appropriate message templates.

Clearly, if computers are to realize their potential as a powerful tool for facilitating learning through reflection, we must identify models for effective re-

*The research described in this paper was supported by the Office of Naval Research, Cognitive and Neural Sciences Division, and a National Science Foundation Research Initiation Award.

flective interactions. Our research is aimed at identifying strategies for choosing what information to include in in responses to students' questions during RFU and for organizing and presenting that information in a manner that is intelligible to students. In order to develop such strategies, we studied human-human reflective interactions in the SHERLOCK domain, and compared these to the human-computer interactions. Reflecting on the differences between these interactions enabled us to identify features of the human-human interaction that we believe are critical for effective tutoring in complex domains.

The Protocol Study

To identify the strategies that human tutors use when participating in reflective dialogues, we collected protocols of tutors interacting with students in post-problem RFU sessions. For each protocol, the student solved a troubleshooting problem using SHERLOCK, and engaged in an RFU session to review his or her problem-solving. To collect the protocols, the system was used to replay each step of the student's solution. After a step is replayed, the human tutor critiques it by marking the action as "good" (<+>) or as "could be improved" (<->)). During our experiments, students were not allowed to view any of the template-based explanations that SHERLOCK could provide. Instead, they were instructed to address all of their questions to the human tutor. The student and tutor communicated by writing messages with pad and pencil. They were physically arranged so that they could each view a screen image of the SHERLOCK simulation, but they were prevented from communicating in any way other than writing messages on the pad. Because SHERLOCK keeps a records of all student actions for each problem session, the student traces can be replayed at any time. After each RFU protocol was gathered, we replayed the trace of the student's actions and collected the messages that SHERLOCK would have produced.

To date, we have collected data from 24 student-tutor interactions with 14 different students and 3 different tutors. This corpus contains approximately 1725 sentences in approximately 232 question/answer pairs. We have analyzed the protocol data, and have identified several features of human expert explanation that are lacking in the template-based approach currently employed in SHERLOCK.

Critical Features of Human Discourse

We found two striking differences between the explanations produced by the human tutor and those produced by SHERLOCK. First, human tutors freely refer to the previous dialogue in their subsequent explanations. This facilitates understanding and learning by relating new information to recently conveyed material, and avoiding repetition of old material that would distract the student from what is new.

Second, human tutors make extensive use of discourse markers. These markers express relationships among individual units of information, thus adding structure to complex explanations and making them easier to understand. Such rhetorical devices affect text cohesion, and research in reading comprehension shows that these devices increase the learner's ability to construct a coherent mental representation of the incoming information, e.g., (Brewer, 1980; Goldman and Durán, 1988; Meyer, Brandt and Bluth, 1980).

Referring to Previous Discourse

In the protocol study, we found that the human explanations were affected by the context created by prior discourse. For example, when students asked follow-up questions, human tutors interpreted and answered these questions in the context of their previous explanations. Clarifying and elaborating on prior explanations requires explainers to understand what they have said previously in order to provide additional, corrective information, and to avoid repeating information that has already been conveyed. Furthermore, even when answering questions that were not follow-up questions, human tutors frequently referred to a previous explanation, e.g., in order to point out similarities (or differences) between the material currently being explained and material presented in earlier explanation(s).

The computer-generated utterances, which do not draw on previous discourse, seem awkward and unnatural. A more serious problem is that in cases where students performed two or more actions that indicated some of the same misconceptions, SHERLOCK simply generated the same comments over and over again. In addition to being frustrating, the system missed out on opportunities to help students form useful abstractions by pointing out the reasons why several specific actions were all symptomatic of the same general type of error.

As an example, consider the dialogue with Sherlock shown in Figure 1. This figure shows the tutor's critique of the action of testing pin 38, followed by the action of testing pin 28. Note several things from this example. First, although these two actions are suboptimal for one of the same reasons, ACTION: You made a non-informative test on the RAG's A1A3A15.

VDC test from pin 38 to ground with a meter. The reading was 28.0000 VDC.

SHERLOCK'S COMMENTS

Student tested off the active circuit path.

Student did not completely test main data signals to the active relay. It is generally more efficient to completely test the main data signals to the active relay.

You tested the data flow input low side, but not the high side.

ACTION: You attempted to verify the secondary data signals affecting the signal path for the RAG's A1A3A15. You made the following measurement:

VDC test from pin 28 to ground with a meter. The reading was 28.0000 VDC.

SHERLOCK'S COMMENTS

Student did not completely test main data signals to the active relay. It is generally more efficient to completely test the main data signals to the active relay.

You tested the data flow input low side, but not the high side.

Figure 1: Sherlock's Comments on Student Actions

TUTOR	<->You tested pin 38 of the A1A3A15.	[1]
STUDENT	Why is testing pin 38 not considered a good step?	[2]
TUTOR	For two reasons. First, because you did not completely test the inputs needed to set relay B24 (you tested the low input but not the high input). These two inputs are critical for the TO test step and are highly suspect when a relay appears to be reset. Secondly, the data input at pin 38 has no effect on the path that is being created between pins 57 & 58 and the output at 11 & 12. Relay B26 could be set or reset – it just doesn't matter.	[3]
	⋮	
TUTOR	<-> You tested pin 28 of the A1A3A15.	[4]
STUDENT	Why is testing pin 28 not a good step?	[5]
TUTOR	*As explained before*, the main data inputs to relay B24 are highly suspect at this time. You have tested the low input to B24 but not the high one. Since it is more likely that there may be a problem with the main data signals to relay B24 than with the secondary data signals to relays B25-B30, the main data signal should be tested first. Testing pin 28 is not a bad step, it is just more likely (on the average) that the main data signals are causing the problem.	[6]
	⋮	
TUTOR	<->You tested pin 36 of the A1A3A15.	[7]
STUDENT	Don't I need to test pin 36?	[8]
TUTOR	You might have to, but *for the same reasons given when you tested pin 28, it is generally more efficient to test the main control data signals first, and then test the secondary control data signals if necessary.*	[9]

Figure 2: Human Tutor's Comments on Student Actions

this is not explicitly stated in the texts. (For expository purposes, italics are used to highlight the similar portions of the explanations in Figure 1.) The student must perform a detailed comparison of the two texts to determine whether and in what ways the two situations are similar and in what ways they differ.

Contrast SHERLOCK's explanations with those produced by a human tutor in Figure 2. This figure shows the human tutor's explanations for 3 actions, the first two of which are the same as those critiqued in the SHERLOCK-generated example above. The human tutor's explanations make it clear that testing pin 28 is bad for one of the same reasons as testing pin 38, and testing pin 36 is bad for precisely the same reason as testing pin 28. In Figure 2, italics are used to highlight what we categorize as contextual effects on the explanations given. For example, when explaining why testing pin 28 is bad (turn 6), the tutor refers back to one of the reasons given in the explanation in turn 3, and reiterates the fact that the main data inputs are highly suspect and have not been completely tested (signalled by "As explained before"). The tutor then introduces the notions of main and secondary data control signals and justifies why the main data signal should be tested first. Later, when explaining why testing pin 36 is bad in turn 9, the tutor refers back to the explanation given when assessing the test of pin 28 and states a generalization explaining why these two actions are considered suboptimal, i.e., that the main data signals should always be tested before secondary data signals. The tutor expects the student to be able to make use of the explanation given in turn 6 (and therefore turn 3) by indicating that it is relevant to the current situation ("for the same reasons given ..." serves this purpose). Accordingly, the tutor does not repeat the detailed explanation of why the main control data signals are suspect, nor why they should be tested first. By generating the explanation in turn 9 in such a way that it meshes with the previous two, not only does the tutor correct the student's error, but forces the student to consider how the three situations are similar. Pointing out this similarity may facilitate the student in forming the domain generalization and recognizing how the three instances fit this generalization.

Based on our study of human-human reflective dialogues, we are developing a taxonomy that classifies the types of contextual effects that occur in our data according to the explanatory functions they serve. Thus far, we have identified four main categories:

- explicit reference to a previous explanation (or portion thereof) in order to point out similarities (differences) between the material currently being explained and material presented in earlier explanation(s),
- omission of previously explained material to avoid distracting the student from what is new,
- explicit marking of repeated material to distinguish it from new material (e.g., "As I said before, ...")
- elaboration of previous material in the form of generalizations, more detail, or justifications.[1]

We are also performing a more detailed study of the corpus in order to determine the conditions under which human tutors refer to previous explanations. In RFU interactions the most commonly asked question is a request to justify the tutor's assessment of a student action (42% of all questions asked during RFU). We found that 27% of the answers to such questions involved references to previous justifications of assessments in order to point out similarities or differences. However, it is important to note that not all justifications of assessment provide opportunities for referring to previous explanations. In order to estimate the percentage of cases in which human explainers referred to previous justifications when an appropriate opportunity arose, we devised a case-based reasoning (CBR) algorithm[2] to find relevant prior justifications. The algorithm computes similarity of student actions based on a set of features that were derived from a cognitive task analysis aimed at identifying the factors that expert avionics tutors use in assessing student's troubleshooting actions (Pokorny and Gott, 1990). We found that human tutors explicitly referred to a prior justification (as in Figure 2) in 73% of the cases identified by the CBR algorithm. Therefore, human explainers refer to previous explanations in the vast majority of the cases where it makes sense to do so, at least when answering this type of question.

Use of Discourse Markers

The second distinguishing feature of human tutorial explanations is the extensive use of discourse markers. As an illustration, consider the two explanations appearing in Figures 3 and 4. The explanation appearing in Figure 3 was produced by SHERLOCK, whereas the one appearing in Figure 4 was produced by a human tutor. Note that Sherlock's explanation is difficult to understand because it does not indicate how the parts of the text are related to one another. For example, SHERLOCK's explanation does not make it clear that the material in 3 elaborates 2 by citing a general principle about troubleshooting, nor that 2 and 3 together provide evidence for the tutor's assessment of the student's step as bad. In addition, 4 provides an additional, independent piece of evidence for why the student's action is considered bad. Next, 5 elaborates to explain how the student can find out more about the status of components. Finally, 6 is a concession indicating that the student's action was correct in one way (a voltage test was appropriate at this location in the circuit.) It is difficult to understand 6 when it appears, because the concession relationship between it and the text in 2-5 (the tutor's evidence supporting the claim that the student's action is

[1]This category breaks up into a number of subcategories in our taxonomy.

[2]See (Rosenblum and Moore, 1993) for details.

ACTION: <-> VDC test from pin 33 to ground on A1A3A8.

SHERLOCK'S COMMENTS ON YOUR SYSTEM UNDERSTANDING:

[1] Student space-split between the UUT, the stimulus and the measurement areas (or between the UUT and the measurement area, if there is no stimulus) before testing the measurement signal path.

SHERLOCK'S COMMENTS ON YOUR STRATEGIC SKILL:

[2] Student tests data before input/output signals.

[3] An efficient testing strategy is to verify that there is a problem on a component's signal path before investigating the component's control data signals. If the signal going through the component is good, then the control data signals are also good.

[4] Student tests pins of unverified component which have been verified by prior TO test.

[5] By clicking on a component on the circuit diagram, Sherlock will tell you what parts of a component are not verified for each troubleshooting step.

[6] Student performs a correct type of test.

Figure 3: Sherlock's comments on student action

⋮

TUTOR	<-> VDC test from pin 33 to ground on A1A3A8.	[1]
STUDENT	Why is testing pin 33 considered a bad move?	[2]
TUTOR	**For several reasons. First,** although you know that the UUT is good, you should eliminate the test package before troubleshooting inside the test station. **This is because** the test package is moved frequently and is **thus** more susceptible to damage than the test station. **Also,** it is more work to open up the test station for testing and the process of opening drawers and extending cards may induce problems which did not already exist. **Second,** it is usually a better strategy to locate a problem along the signal flow path before suspecting that the data signals are causing the fail. You really should test the signal flow input and output signals first and then decide if testing the data flow signals is necessary. **Finally, since** TO test 2 passed, you should already know that the input on pin 33 is probably good (TO test 2 used TPA63 which needs the same input on 33). **Therefore,** testing pin 33 is really a redundant move.	[3]

Figure 4: Human tutor's critique of student action

bad) is not signalled.

Contrast this with the human tutor's explanation, which clearly states that there are several reasons why the student's action was assessed negatively. In explaining each of the reasons, the human tutor supplies justification for the claim that the student's action can be considered bad. Note that the human tutor's explanation includes many discourse markers that convey important relationships between the information that appears in the text. (These appear in bold type in Figure 4.) For example, the tutor signals evidence for a claim with markers such as "because" and "since". When he argues from evidence to claim, as in "the test package is moved frequently and is thus more susceptible to damage ...", he uses markers such as "thus" and "therefore" to indicate the claim. Finally, he indicates where the argument for each reason starts and ends with the markers "First," "Second," and "Finally." The marker "also" is used to indicate additional justification within a reason. These markers make explicit the intentional and informational (semantic) relationships between the parts of this complex text, and thus make it easier to understand.

The problem of determining when discourse markers should be used, and which markers would be most effective in increasing the student's comprehension of the explanation is an open research problem. To tackle this problem, we have begun a detailed linguistic analysis of the explanations in our corpus. From a pilot study, we have reason to hypothesize that marker selection is influenced by the intentional and informational relations (Moore and Pollack, 1992) between text segments, the topic structure of the text, the size of the segments being related, and the embedding of relations in the hierarchical structure of the text. From this study, we expect to develop a catalogue of the discourse markers used in explanations in the SHERLOCK domain and the features that predict usage of each marker. This information will then be used to construct a computational model that will enable our explanation generator to select appropriate discourse markers.

Conclusions

Comparison of human explanations with those produced by a computer system using a template-based explanation process has enabled us to identify two properties of human discourse that seem crucial for producing effective explanations in reflective interactions. The next step is to build computational systems that are capable of producing explanations that have these properties. We have already made progress toward building a system that takes prior utterances into account when planning explanations. In (Carenini and Moore, 1993; Rosenblum and Moore, 1993), we describe the strategies we have implemented for identifying relevant prior explanations, and the mechanisms that enable our explanation planner to exploit the information stored in its discourse history in order to omit information that has previously been communicated, to point out similarities and differences between entities and situations, and to mark re-explanations in circumstances where they are deemed appropriate. In future work, we will implement strategies for selecting discourse markers to convey the relationships between units of information in complex texts.

In order to evaluate the effectiveness of the properties we have identified, we are designing our explanation facility so that the abilities to integrate previous explanations into current explanations, and to employ discourse markers, are optional facilities that can be enabled or disabled. Thus we will be able to systematically evaluate the effect of these two capabilities on students' satisfaction with the system, their comprehension of explanations, and their learning of complex problem-solving strategies.

References

Brewer, W. F., 1980. Literacy theory, rhetoric, and stylistics: Implications for psychology. In Shapiro, R. J. *et al.* (Eds.), *Theoretical Issues in Reading Comprehension*, 221–239. Hillsdale, NJ: Lawrence Erlbaum.

Carenini, G. and Moore, J. D., 1993. Generating explanations in context. In Gray, W. D. *et al.* (Eds.), *Proceedings of the International Workshop on Intelligent User Interfaces*, 175–182. Orlando, Florida: ACM Press.

Collins, A. and Brown, J. S., 1988. The computer as a tool for learning through reflection. In Mandl, H. *et al.* (Eds.), *Learning Issues for Intelligent Tutoring Systems*, 1–18. NY: Springer-Verlag.

Goldman, S. and Durán, R. P., 1988. Answering questions from oceanography texts: Learner, task, and text characteristics. *Discourse Processes* 1:373–412.

Gott, S. P., 1989. Apprenticeship instruction for real world tasks: The coordination of procedures, mental models, and strategies. In Rothkopf, E. Z. (Ed.), *Review of Research in Education*, volume XV, 97–169.

Lesgold, A., Lajoie, S., Bunzo, M., and Eggan, G., 1992. Sherlock: A coached practice environment for an electronics troubleshooting job. In *Computer Assisted Instruction and Intelligent Tutoring Systems: Shared Goals and Complementary Approaches*, 201–238. Hillsdale, New Jersey: Lawrence Erlbaum.

Lesgold, A., in press. Assessment of intelligent training technology. In Baker, E. L. and O'Neil, H. F., (Eds.), *Technology Assessment*. Hillsdale, NJ: Lawrence Erlbaum.

Meyer, B. J. F., Brandt, D. M., and Bluth, G. J., 1980. Use of top-level structure in texts: Key for reading comprehension in ninth-grade students. *Reading Research Quarterly* 16:72–102.

Moore, J. D. and Pollack, M. E., 1992. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics* 18(4).

Owen, E. and Sweller, J., 1985. What do students learn while solving mathematics problems? *Journal of Educational Psychology* 77:272–284.

Pokorny, R. and Gott, S., 1990. The evaluation of a real-world instructional system: Using technical experts as raters. Technical report, Armstrong Laboratories, Brooks Air Force Base.

Rosenblum, J. A. and Moore, J. D., 1993. Participating in instructional dialogues: Finding and exploiting relevant prior explanations. In *Proceedings of the World Conference on Artificial Intelligence in Education*.

Sweller, J., 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12:257–285.