My Science Tutor: Learning Science with a Conversational Virtual Tutor

Sameer Pradhan   Ron Cole   Wayne Ward
Boulder Learning, Inc.
Boulder, CO
{pradhan,rcole,wward}@boulderlearning.com

Abstract

This paper presents a conversational, multimedia, virtual science tutor for elementary school students. It is built using state of the art speech recognition and spoken language understanding technology. This virtual science tutor is unique in that it elicits self-explanations from students for various science phenomena by engaging them in spoken dialogs and guided by illustrations, animations and interactive simulations. There is a lot of evidence that self-explanation works well as a tutorial paradigm, Summative evaluations indicate that students are highly engaged in the tutoring sessions, and achieve learning outcomes equivalent to expert human tutors. Tutorials are developed through a process of recording and annotating data from sessions with students, and then updating tutor models. It enthusiastically supported by students and teachers. Teachers report that it is feasible to integrate into their curriculum.

1 Introduction

According to the 2009 National Assessment of Educational Progress (NAEP, 2009), only 34 percent of fourth-graders, 30 percent of eighth-graders, and 21 percent of twelfth-graders tested as proficient in science. Thus, over two thirds of U.S. students are not proficient in science. The vast majority of these students are in low-performing schools that include a high percentage of disadvantaged students from families with low socioeconomic status, which often include English learners with low English language proficiency. Analysis of the NAEP scores in reading, math and science over the past twenty years indicate that this situation is getting worse. For example, the gap between English learners and English-only students, which is over one standard deviation lower for English learners, has increased rather than decreased over the past 20 years. Moreover, science instruction is often underemphasized in U.S. schools, with reading and math being stressed.

The Program for International Student Assessment (PISA), coordinated by the Organization for Economic Cooperation and Development (OECD), is administered every three years in 65 countries across the world. According to their findings in 2012, the U.S. average science score was not measurably different from the OECD average.

Our approach to address this problem is a conversational multimedia virtual tutor for elementary school science. The operating principles for the tutor are grounded on research from education and cognitive science where it has been shown that eliciting self-explanations plays an important role (Chi et al., 1989; Chi et al., 1994; Chi et al., 2001; Hausmann and VanLehn, 2007a; Hausmann and VanLehn, 2007b). Speech, language and character animation technologies play a central role because the focus of the system is on engagement and spoken explanations by students during spoken dialogs with the virtual tutor. Summative evaluations indicate that students are highly engaged in the tutoring sessions, and achieve learning outcomes equivalent to expert human tutors (Ward et al., 2011; Ward et al., 2013). Surveys of participating teachers indicate that it is feasible to incorporate the intervention into their curriculum. Also, importantly, most student surveys indicate enthusiastic support for the system.

Tutorials are developed through an iterative process of recording, annotating and analyzing logs from sessions with students, and then updating tutor models. This approach has been used to de-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics—System Demonstrations, pages 121–126,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

velop over 100 tutorial dialog sessions, of about 15 minutes each, in 8 areas of elementary school science.

My Science Tutor (MyST) provides a supplement to normal classroom science instruction that immerses students in a multimedia environment with a virtual science tutor that models an engaging and effective human tutor. The focus of the program is to improve each student's engagement, motivation and learning by helping them learn to visualize, reason about and explain science during conversations with the virtual tutor. The learning principles embedded in MyST are consistent with conclusions and recommendations of the National Research Council Report, "Taking Science to School: Learning and Teaching Science in Grades K-8" (NRC, 2007), which emphasizes the critical importance of scientific discourse in K-12 science education. The report identifies the following crucial principles of scientific proficiency:

Students who are proficient in science:

1. *Know, use, and interpret* scientific explanations of the natural world;
2. *Generate and evaluate* scientific evidence and explanations;
3. *Understand* the nature and development of scientific knowledge; and
4. *Participate productively* in scientific practices and discourse.

The report also emphasizes that scientific inquiry and discourse is a learned skill, so students need to be involved in activities in which they learn appropriate norms and language for productive participation in scientific discourse and argumentation.

2 The MyST Application

MyST provides students with the scaffolding, modeling and practice they need to learn to reason and talk about science. Students learn science through natural spoken dialogs with the virtual tutor Marni, a 3-D computer character. Marni asks students open-ended questions related to illustrations, silent animations or interactive simulations displayed on the computer screen.

Figure 1 shows the student's screen with Marni asking questions about media displayed in a tutorial. The student's computer shows a full screen window that contains Marni, a display area for presenting media and a display button that indicates the listening status of the system. Marni produces accurate visual speech, with head and face movements that are synchronized with her speech. The media facilitate dialogs with Marni by helping students visualize the science they are discussing. The primary focus of dialogs is to elicit explanations from students. MyST compares the student's spoken explanations to reference explanations for the lesson by matching the extracted *semantic roles* using the Phoenix parser (Ward, 1991), then presents follow-on questions and media, to help the student construct a correct explanation of the phenomena being studied. The virtual tutor Marni, who speaks with a recorded human voice, is designed to model an effective human tutor that the student can relate to and work with to learn science. MyST provides a non-threatening and supportive environment for students to express their ideas. The dialogs scaffold learning by providing students with support when needed until they can apply new skills and knowledge independently.

MyST is intended to be used as an intervention for struggling students, with intended users being K-12 science students. While it should prove a benefit to all students, struggling students should benefit most. Depending on the recording conditions and ambient noise, as well as the characteristics of the student and session, the recognition word error rate ranges from low 20s to mid-40s. MyST will contain tutorials for 3 topics per grade, with content aligned with NGSS. For each topic, students engage in an average of 10 spoken dialog sessions with the tutor, lasting approximately 20 minutes each. oThe MyST tutorial sessions are in addition to the normal classroom instruction for the module. Tutoring sessions can be assigned as homework or during regular school hours, at the teacher's discretion. In the initial studies, tutoring was always done during regular school hours. Teachers specify the space in the school to be used, generally any relatively quiet room. Students are sent to use the system a few at a time, depending on how many computers are available (5 computers per classroom were used in the efficacy study). All students are given a demo at the beginning of the school year and given a chance to ask questions. Teachers schedule time for students, but students log on and use the system without supervi-

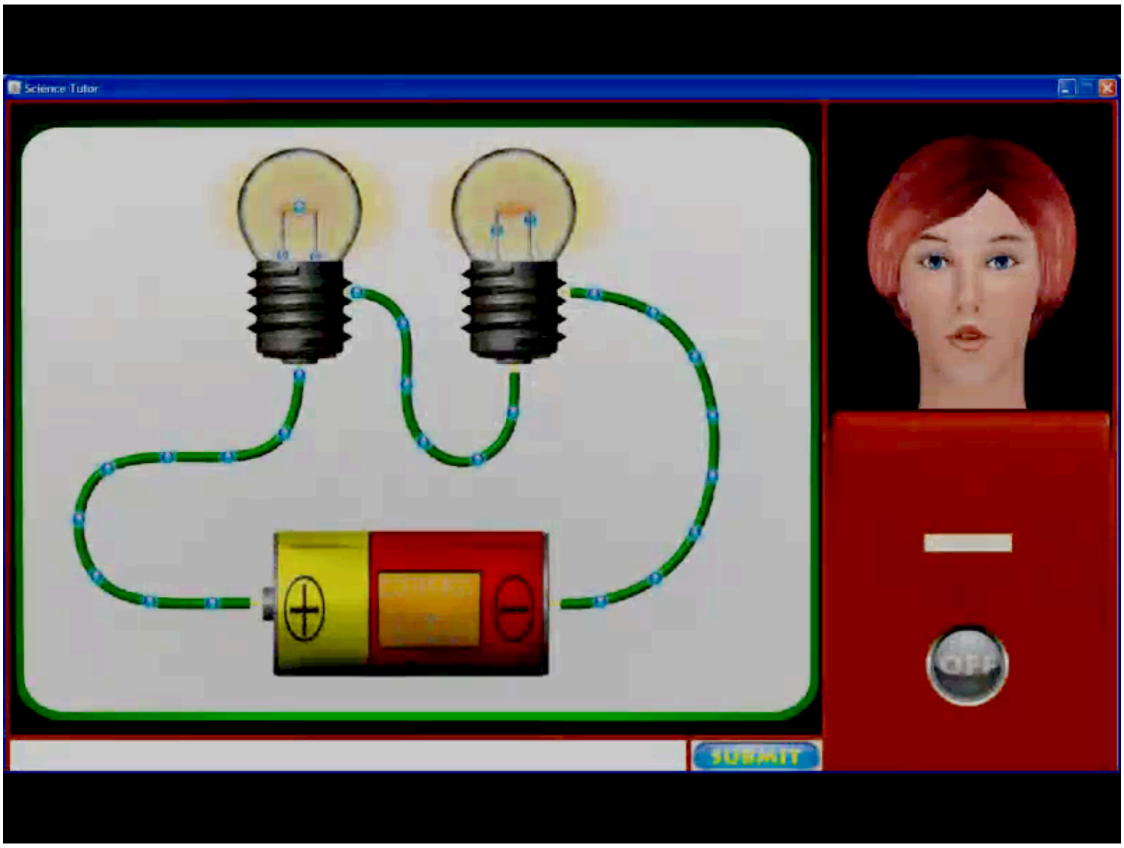

Figure 1: A snapshot of the screen as seen by a student.

sion, so it has minimal impact on teacher time or other human resources. In studies thus far, surveys report that teachers did not have problems using the system and it did not interfere with their other activities.

The application will eventually be deployed using a Software as a Service (SaaS) model. It will run on a server and students will access it through their browser. If internet service is not available or reliable, it can be run stand-alone and the data uploaded when service is available. Both content and user populations will evolve and system models need to incorporate dynamic adaptation in an efficient way. Data from all user sessions is logged in a database and is available for continuous evaluation and re-training of system models. The system is designed to work well even if it doesn't understand the user, but becomes more engaging and efficient as it understands the user better. As training data grows model parameters become more accurate and more explicit models are trained, such as acoustic models for ELL students. Unsupervised training is combined with active learning to optimize use of the data for tuning system models. Teachers in the initial studies did not feel that they would have a problem implementing the system.

3 Theoretical Framework

The theory of change, and theoretical and empirical support Science curricula are structured with new concepts building on those already encountered. Struggling students fall further and further behind if they don't understand the content of each topic. Research has demonstrated that human tutors are effective (Bloom, 1984; Madden and Slavin, 1989), media presentations are effective (Mayer, 2001) and QtA dialog strategies are effective (Murphy and Edwards, 2005). A system that emulates a human tutor using media presentations to focus a student's attention and conducting a QtA-style dialog with the student should also be effective. This additional time spent thinking and talking about the science concepts covered in class will enable students who would have fallen behind to understand the content of the current investigation so they will be prepared to participate in and understand subsequent topics. Student learning will increase because they are excited about and engaged by interesting and informative presentations that help them visualize and understand the science and because they will learn to engage in conversations in which they construct, reflect on and revise mental models and explanations about the science they are seeing and trying to explain. MyST dialogs are designed to provide students with understandable multimedia scenarios, explanations and challenges and a supportive social context for communication and learning. Science is introduced through scenarios that students can relate to and make sense of, and provide a context for introducing and using science vocabulary and making connections between vocabulary, objects, concepts and their prior knowledge. Multimedia learning tools show and explain science, and then enable students to revisit the media and explain the science in their own words.

Research has demonstrated that having students produce explanations improves learning (Chi et al., 1989; Chi et al., 2001; King, 1994; King et al., 1988; Palincsar and Brown, 1984). In a series of studies, Chi et al. (1989; 2001) found that having college students generate self-explanations of their understanding of physics problems improved learning. Self-explanation also improved learning about the circulatory system by eighth grade students in a controlled experiment (Chi et al., 1994). Hausmann and Van Lehn (2007a; 2007b) note that: "self-explaining has consistently been shown to be effective in producing robust learning gains in the laboratory and in the classroom." Experiments by Hausmann and Van Lehn (Hausmann and VanLehn, 2007a) indicate that it is the process of actively producing explanations, rather than the accuracy of the explanations, that makes the biggest contribution to learning.

4 Semantic Underpinnings

The patterns used in MyST to extract frames from student responses are trained from annotated data. The specification of tutorial semantics begins with creating a narrative. A tutorial narrative is a set of natural language statements that express the concepts to be discussed in as simple a form as possible. These do not represent the questions that the system asks, but are the set of points that the student should express.

The narrative represents what an ideal explanation from a student would look like. The narrative statements are manually annotated to reflect the desired semantic parses. These parsed statements define the domain of the tutorial. The initial grammar patterns are extracted from the narratives and have all of the roles and entities that will be discussed, but only a few (or one) ways of expressing them. As the system is used, the grammar is expanded to cover the various ways students articulate their understandings of the science concepts. This is done by annotating recordings of student responses generated in real use. So the life cycle of the natural language processing model for a module is:

1. Create and annotate a narrative to define the domain of the tutorial
2. Field the system to collect data from real users
3. Sample incoming data and annotate
4. Evaluate current model and re-train
5. Repeat step 3-4 as long as the module is used

As the system is used, it logs all transactions and records student speech. When tutorials are deployed for live use, incoming data are processed automatically to assess system confidence in the interpretation of student responses. High-confidence items are added to the training database, and low confidence sessions are selected for transcription and annotation. The system also provides a text input mode that students can use to interact with the Avatar. Once annotated, the data are added to the training set and system models (acoustic models, language models and extraction patterns) are retrained. Periodically, data are sampled for test sets and a learning curve is plotted for each module. All elements of this process are automatic except for transcription and annotation.

The semantics of each domain are constrained, but student responses can vary greatly in the ways they choose to express concepts and terms. It takes time, effort and data to get good coverage of student responses. Semantic annotation for the system consists of annotating:

Entities—*The basic concepts talked about in the session and the phrases that would be considered synonyms.* Electricity could be expressed as electricity, energy, power, current or electrical energy. Coverage of term synonyms from annotated data is generally achieved fairly quickly. **Roles**—*How the entities in an event or concept are related* *to each other.* The larger problem is to attain coverage of the patterns discriminating between possible role assignments. Not only is there more disfluency and variability here, annotating them is a more difficult task for someone not trained to do it. Currently, it takes about one hour for a highly-trained annotator to mark up the data collected in a single 20-minute tutorial session.

5 Extrinsic Evaluation

An assessment was conducted in schools to compare learning gains from human tutoring and MyST tutoring to business-as-usual classrooms. Learning gain was measured using standardized assessments given to students in each condition before and after each science module. Both tutoring conditions had significantly higher learning gains than the control group. While the effect size for human tutors vs. control (d=0.68) was larger than for MyST vs. control (d=0.53), statistical tests supported the hypothesis of no significant difference between the two.

A simple two-group comparison using a Repeated Measures ANOVA shows a statistically significant effect at F=46.4, df 1,759, p <.0001 favoring the treatment group. The interaction between group and module was also significant at F=9.5, p < .001. We also used an Analysis of Covariance (ANCOVA) to compare post-test scores. This procedure adjusts for pre-test differences while comparing the post-test average scores. The two-group comparison was significant at F=7.4, df 1,768, p=.018. We also saw a significant interaction between treatment group and module with F=12.4, df 3,768. Testing the main effects with a hierarchical mixed model with students nested within classrooms we found a significant effect for the treatment group at F=6.2, df 1,2l7,662, p=0.013. No significant interaction effect was found for module by group.

A written survey was given to the students who participated in the gas. Measures were taken to avoid bias wherein students give overly positive answers to questionnaires. The survey included questions that asked for ratings of student experience and impressions of the program and its usability. Across schools, 47% of students said they would like to talk with Marni after every science investigation, 62% said they enjoyed working with Marni “a lot,” and 53% selected “I am more excited about science” after using the program. Only 4% felt that the tutoring did not help. Teachers were asked for anonymous feedback to help assess the feasibility of an intervention using the system and their perceptions of the impact of the system. A teacher survey was given to all participating teachers directly after their students completed tutoring. The survey asked teachers about the perceived impact of using Marni for student learning and engagement, impacts on instruction and scheduling, willingness to potentially adopt Marni as part of classroom instruction, and overall favorability toward participating in the research project. Teachers answered items related to potential barriers in implementing new technology in the classroom. 100% of responding teachers said that they felt it had a positive impact on their students, they would be interested in the program if it were available and they would recommend it to other teachers. 93% said that they would like to participate in the project again. 74% indicated that they would like to have all of their students use the system (not just struggling students). Following these studies, Boulder Learning combined the best elements of the initial systems into the current MyST system, and with continued funding from IES (Cognition and Student Learning Goal 3), is conducting an efficacy study. We are currently in the 3rd year of a 4 year study. While data collection will continue for another year, preliminary results support the learning gain performance from the initial studies.

6 MyST Conversations Corpus of Student Speech (MCCSC)

We are making a cleaned up version of the corpus available to the research community[1] for free and for commercial use at a pre-determined cost. The first release of the corpus v0.1.0 comprises 298 hours of speech out of which 198 hours are manually transcribed. This covers roughly 1.4 million words of text. We are in the process of cleaning up about the same amount of collected data for future distribution.

7 Future Work

In the near future we plan to evaluate applying a statistical labeler trained on existing corpora to the task of Role assignment. This approach should provide increased robustness to novel input and substantially reduce the human annotation effort required to attain a given level of coverage. The

[1] `http://corpora.boulderlearning.com/myst`

Proposition Bank (PropBank) provides a corpus of sentences annotated with domain-independent semantic roles (Palmer et al., 2005). PropBank has been widely used for the development of machine learning based Semantic Role Labeling (SRL) systems. Pradhan et al. (2005) used a rich set of syntactic and semantic features to obtain a performance with F-score in the low-80s. It has been an integral component of most question answering systems for the past decade. Since its first application to the newswire text, PropBank has been extended to cover many more predicates and diverse genres in the DARPA OntoNotes project (Weischedel et al., 2011; Pradhan et al., 2013) and the DARPA BOLT program. We plan to map PropBank SRL output onto MyST frames. Domain specific entity patterns will still need to be applied to produce the canonical extracted form, but that is a much simpler task than role assignment and one more suited to non-linguists.

References

B. Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16.

M. Chi, M. Bassok, M. Lewis, P. Reimann, R. Glaser, and Alexander. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2).

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

M. T. H. Chi, S. A. Siler, H. Jeong, T. Yamauchi, and R. G. Hausmann. 2001. Learning from human tutoring. *Cognitive Science*, 25(4):471–533.

R. G. M. Hausmann and K. VanLehn. 2007a. Explaining self-explaining: A contrast between content and generation. *Artificial Intelligence in Education*, pages 417–424.

R. G. M. Hausmann and K. VanLehn. 2007b. Self-explaining in the classroom: Learning curve evidence. In *29th Annual Conference of the Cognitive Science Society*, Mahwah, NJ.

A. King, A. Staffieri, and A. Adelgais. 1988. Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *Journal of Educational Psychology*, 90(1):134–152.

A. King. 1994. Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal*, 31(2).

N. A. Madden and R. E. Slavin. 1989. Effective programs for students at risk. In R. E. Slavin, N. L. Karweit, and N. A. Madden, editors, *Effective pull-out programs for students at risk*. Allyn and Bacon.

R. Mayer. 2001. *Multimedia Learning*. Cambridge University Press., Cambridge, U.K.

P. K. Murphy and M. N.b Edwards. 2005. What the studies tell us: A meta-analysis of discussion approaches. In *American Educational Research Association*, Montreal, Canada.

National Research Council. NRC. 2007. Taking science to school: Learning and teaching science in grades k-8. In R. A. Duschl, H. A. Schweingruber, and A. W. Shouse, editors, *Committee on Science Learning Kindergarten through Eighth Grade. Washington D.C.* The National Academies Press.

A. Palincsar and A. Brown. 1984. Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1(2).

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Sameer Pradhan, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James Martin, and Dan Jurafsky. 2005. Support vector learning for semantic argument classification. *Machine Learning*, 60(1):11–39.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria, August.

W. Ward, R. Cole, D. Bolanos, C. Buchenroth-Martin, E. Svirsky, S. V. Vuuren, and L. Becker. 2011. My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Trans. Speech Lang. Process.*, 7(4).

Wayne Ward, Ron Cole, Daniel Bolanos, C. Buchenroth-Martin, E. Svirsky, and Tim Weston. 2013. My science tutor: A conversational multimedia virtual tutor. *Journal of Educational Psychology*, 105(4):1115–1125.

W Ward. 1991. Understanding spontaneous speech: the phoenix system. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 365–367 vol.1, April.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A large training corpus for enhanced processing. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.