An Evaluation of a Hybrid Language Understanding Approach for Robust Selection of Tutoring Goals

Carolyn Rosé, *Carnegie Mellon University, Language Technologies and Human-Computer Interaction, 5000 Forbes Avenue, Pittsburgh PA, 15213, USA*
cprose@cs.cmu.edu
http://www.cs.cmu.edu/~cprose

Kurt VanLehn, *University of Pittsburgh, Learning Research and Development Center, 3939 O'Hara Street, Pittsburgh PA, 15260, USA*
vanlehn@pitt.edu
http://www.pitt.edu/~vanlehn

Abstract. In this paper, we explore the problem of selecting appropriate interventions for students based on an analysis of their interactions with a tutoring system. In the context of the WHY2 conceptual physics tutoring system, we describe CarmelTC, a hybrid symbolic/statistical approach for analysing conceptual physics explanations in order to determine which Knowledge Construction Dialogues (KCDs) students need for the purpose of encouraging them to include important points that are missing. We briefly describe our tutoring approach. We then present a model that demonstrates a general problem with selecting interventions based on an analysis of student performance in circumstances where there is uncertainty with the interpretation, such as with speech or text based natural language input, complex and error prone mathematical or other formal language input, graphical input (i.e., diagrams, etc.), or gestures. In particular, when student performance completeness is high, intervention selection accuracy is more sensitive to analysis accuracy, and increasingly so as performance completeness increases. In light of this model, we have evaluated our CarmelTC approach and have demonstrated that it performs favourably in comparison with the widely used LSA approach, a Naive Bayes approach, and finally a purely symbolic approach.

Keywords. Tutorial dialogue, language understanding, evaluation

INTRODUCTION

In this article we address technical challenges at the heart of the problem of supporting the development of explanation skills for articulating a correct conceptual understanding of physics. We present both a technological approach as well as a framework in which to evaluate

technology for addressing this and similar problems in terms of the impact the performance of the underlying technology has on the experience of the student interacting with the technology.

Recent studies of human tutoring suggest that a productive activity for teaching is to have students explain physical systems qualitatively (Chi et al, 1994). While students in elementary mechanics courses have demonstrated an ability to master the skills required to solve quantitative physics problems, a number of studies have revealed that the same students perform very poorly when faced with qualitative physics problems (Halloun & Hestenes, 1985; Hake, 1998). Furthermore, the naïve conceptions of physics that they bring with them when they begin a formal study of physics do not change significantly by the time they finish their classes (Halloun & Hestenes, 1985). In our previous work on the Atlas-Andes system, we have demonstrated one way an implemented tutorial dialogue system can address this problem (Rosé et al., 2001a). We evaluated two versions of the Andes coached problem solving environment (Gertner & VanLehn, 2000), one using tutorial dialogue in addition to hints and one with hints only. Our evaluation demonstrated that the inclusion of tutorial dialogue technology as conceptual support improves student conceptual understanding of physics while not significantly increasing the total amount of time students spend with the environment. Building upon this early success, the goal of the WHY2 project (VanLehn et al., 2002) is to focus specifically on conceptual physics problems and support students in developing the skills to articulate multi-step conceptual physics explanations.

The work reported in this article grows out of a long term and multi-faceted effort to emulate in intelligent tutoring technology the elements that make human tutoring such an effective form of instruction. Expert human tutors are highly successful at educating students (Bloom, 1984; Cohen, Kulik & Kulik, 1982). Students working with an expert human tutor achieve a learning gain of up to two standard deviations above those in a regular classroom setting. Emulating this "2 sigma effect" has long been the holy grail of intelligent tutoring research. While great strides in developing instructional technology have been made especially in the area of building coached problem solving practice environments (Gertner & VanLehn, 2000; Koedinger et al., 1997), achieving the goal of the full extent of the effectiveness of expert human tutors remains elusive. The search for the answer to this mystery has taken many forms, but one common thread through generations of investigation has been the belief that the answer lies in the natural language dialogue that is the dominant form of communication between students and human tutors (Carbonell, 1969; Rosé et al., 2001b; Person & Graesser, 2003).

While early efforts to emulate the effectiveness of human tutorial dialogue, such as the SCHOLAR system (Carbonell, 1969; Carbonell, 1970) and the original WHY system (Stevens & Collins, 1977), were landmark systems in the history of intelligent tutoring research, they were naïve both theoretically and technologically. The conception of what Socratic tutoring is and why it should be effective was not sufficiently well defined, and the technology to support such interactions was not yet mature. As a result, the efforts were unsuccessful, and active interest in pursuing the goal of tutorial dialogue systems died out. For more than two decades, the language technologies and intelligent tutoring communities largely went their own ways. One exception in the late 80s was a very rudimentary form of Socratic tutorial dialogue technology deployed as part of the Scientific Reasoning Series (Bork, 2004). During this two decade period of time, a wide range of technologies for supporting learning that do not involve natural language interactions were developed, evaluated, and refined.

With the dawn of the 21st century, when both the fields of intelligent tutoring and language technologies had substantially matured, there was a resurgence of interest in bringing these two communities together once again. Wielding state-of-the-art dialogue management (Freedman, 2000) and language understanding technology (Glass, 1999; Wiemer-Hastings, Wiemer-Hastings & Graesser, 1999; Rosé, 2000; Rosé & Lavie, 2001; Rosé et al., 2002), intelligent tutoring researchers have made great strides in building and evaluating tutorial dialogue systems with students, often in realistic educational settings (Graesser, Bowers, Hacker, & Person, 1998; Rosé et al., 2001a; Heffernan & Koedinger, 2002; Ashley, Desai, & Levine, 2002; Graesser, VanLehn, the TRG, & the NLT, 2002; Aleven, Koedinger, & Popescu, 2003; Evens & Michael, 2004). The work reported in this article is one of these recent proofs of the current technological feasibility of building tutorial dialogue systems. These formative evaluation studies demonstrate that current computational linguistics technology is sufficient for building tutorial dialogue systems that are robust enough to be put in the hands of students and to provide useful learning experiences for them. However, they have not yet yielded the dramatic improvements over more standard forms of tutoring systems that were predicted. These evaluations of recent tutorial dialogue systems demonstrate that the secret to achieving the same effectiveness of expert tutorial dialogue does not necessarily lie simply in a modality shift to natural language input and output.

Nevertheless, we argue that the work presented here constitutes one more step towards unlocking the mystery of the elusive "2 sigma effect". One can easily characterize all types of interactions as dialogues in some sense. In this way, dialogues can be said to vary along three primary dimensions: namely, characteristics of structure, reactivity or flexibility in initiative, and modality. Looking at naturalistic human tutorial dialogue inspires us to broaden our view beyond standard coached problem solving environments, and to consider forms of interaction and roles for the technology to play that are not typically supported by current intelligent tutoring systems. Human tutors play different roles in their tutoring beyond coached problem solving practice and engage in a much wider range of interactions that vary along all three dimensions mentioned above. While principles of effective coached problem solving practice environments are well understood, supporting the development of explanation skills is far less explored and as a result, far less well understood. This article presents an investigation into that role, which happens to be one that naturally relies heavily on natural language interaction. Thus, while it is yet to be demonstrated that the primary benefit in human tutorial dialogue lies specifically in the natural language itself, the capability of making use of state-of-the-art language technology brings new forms of human-computer interaction for educational benefit within the realm of what is practical to pursue and evaluate.

In the remainder of this article we begin by describing our instructional approach and the need for technology capable of analyzing extended student explanations. Next, we describe a framework for evaluating technological solutions to the essay analysis problem. We then describe one technical approach to this problem, specifically a novel hybrid text classification approach. While purely bag-of-words approaches to automatic text classification have been widely and successfully used in the intelligent tutoring community (Graesser et al., 2000; Wiemer-Hastings, 2000), we argue that domain considerations are important to consider in connection with feature selection for text classification approaches. In particular, since physics is a causal domain, we argue that features from a deep syntactic functional analysis of the text are important for accurate classification in this domain. Finally, we present an evaluation of this

technological solution using our proposed framework. Our evaluation demonstrates both the practicality of making use of features from a deep syntactic functional analysis as well as their contribution to the success of the approach. We conclude with discussion of our results and current directions.

INSTRUCTIONAL APPROACH

Some physics educators are recognizing the need to engage students in active and interactive forms of physics instruction (Meltzer & Manivannan, 2002). The purpose is to keep students engaged in scientific reasoning and to actively monitor student understanding in a way that is practical during their lecture time. As a complementary effort, we are developing an environment called WHY2 as an interactive form of instruction available to engage students in a deeper and more extensive suite of reasoning tasks including explanation activities outside of their lecture time. The goal of the WHY2 systems is specifically to support the development of conceptual reasoning and explanation skills by challenging students with qualitative physics questions that require multi-step conceptual reasoning to answer fully.

To illustrate our approach, consider the following, which we refer to as The Pumpkin Problem: "Suppose a man is running in a straight line at constant speed. He throws a pumpkin straight up. Where will it land? Explain." Fully answering this question requires the interaction of multiple crucial physics concepts. For example, one piece of required understanding is the idea of the independence between forces in the horizontal and vertical directions, a foundational piece of knowledge for physics reasoning. Students must also understand that contact forces such as that exerted by the runner while he is still holding the pumpkin end as soon as the contact ends, a notion not always grasped by students even after a semester of college level physics. Building upon that understanding, students must realize that in the absence of a force, velocity remains constant, rather than decreasing as some students believe. Furthermore, students must consider the connection between velocity and displacement. Only if they see this connection can they reason that if the velocity of the pumpkin in the horizontal direction remains constant, then the displacement of the pumpkin and the man from the point of release will always be the same. From this students can infer that when the pumpkin comes down, it will land in the man's hands. Missing any piece along this chain of reasoning would prevent students from fully constructing a satisfying explanation. Some students guess that the pumpkin will land in the man's hands but cannot explain why. Faulty reasoning leads other students to believe the pumpkin will land behind the man.

The ability to construct multi-step lines of reasoning like this requires knowledge and skills at multiple levels. At the most basic level is the understanding of single inference rules in isolation, such as Newton's laws. Just beyond this is the ability to apply an inference rule given a set of preconditions. More advanced than this is selecting an appropriate inference rule given a goal and a set of preconditions. Finally, beyond all these is the ability to plan and execute a multi-step line of reasoning involving the selection and application of a sequence of inference rules. If a student fails to construct a satisfactory explanation in response to a question such as The Pumpkin Problem, the break down in the student's reasoning may have occurred at any place along this continuum. Because our aim is to support students in constructing explanations as

independently as possible, our basic approach is to identify where along this continuum the student is failing and to offer support at the appropriate level. In the remainder of this section, we describe our instructional objectives in greater depth along with support from the science education literature. We then describe the various levels of scaffolding we provide students with in order to tailor our support to the level of the individual student's needs. Next, we illustrate our instructional process with concrete examples. Finally, we offer some motivation for the underlying text-classification approach to student explanation assessment that we explore in the remainder of the article.

Instructional Goals

The ultimate goal of this work is to produce scientific thinkers who see principles of physics at work in their lives. Beyond using those principles to reason about what they see in the world around them, our aim is for them to be able to effectively communicate their reasoning in natural language. The value of supporting the development both of conceptual reasoning skills as well as communication skills as they are manifest in elaborated explanations is well substantiated in the science and mathematics education literature.

The first argument in favor of a qualitative physics emphasis is one of need. Earlier in this article we cited evidence that gaining a conceptual understanding of physics is challenging for students and not an automatic consequence of traditional, quantitatively oriented physics curricula. Students bring naïve misconceptions and conceptual gaps with them into their physics instruction. While some contemporary physics text books (Hewett, 1987) contain more of a conceptual flavor than earlier ones, an unfortunately still too typical pattern for students is that rather than gaining the deep understanding of physics that would allow them to leave their faulty understanding behind, they learn instead to match surface features of problems to equations, and then use a shallow "plug and chug" strategy to obtain a correct solution (Maloney, 1994). While typical coached practice environments are successful at enhancing student problem solving skills, they have similarly been criticized for failing to encourage deep learning (Vanlehn et al., 2000). If students do not reflect upon the hints they are given during problem solving, but instead simply continue guessing until they perform an action that receives positive feedback, they may learn the right actions for the wrong reasons (Aleven et al., 1998; Aleven et al., 1999). One of our goals is to use natural language dialogue and explanation activities to offer students the opportunity to actively reflect upon what they have been taught and to constructively draw conclusions from it. A potential future direction is to combine the reflection and explanation support we develop here with more typical coached problem solving environments, similar to our earlier Atlas-Andes work (Rosé et al., 2001a), in order to offer students a broader range of valuable instructional experiences.

A second argument for emphasizing a qualitative, conceptual understanding of physics is that it has been found to be motivating for students. This has been argued from analyses of data from the Third International Mathematics and Science Study (TIMSS)[1] comparing mathematics and science achievement in American, German, and Japanese schools. These analyses have demonstrated a correlation between student motivation and achievement and the presence of

[1] http://www.mpib-berlin.mpg.de/en/forschung/eub/Projekte/timss.htm

constructivist oriented activities in classrooms that are aimed at fostering qualitative understanding of math and science (Kunter & Baumert, 2004). Learning situations that explicitly connect scientific principles in a tangible way with students' former experiences have been demonstrated to encourage deep engagement with the material and seem to offer valuable opportunities for conceptual growth (Duit & Confrey, 1996). Baumert & Köller (2000) present evidence that classrooms where students work on complex problems without clear-cut solutions and as part of that process are asked to search for connections between various concepts and ideas produced students who achieved higher scores on the TIMSS test. This interpretation of the TIMSS data is further supported by a study by Staub and Stern (2002), which showed that elementary school children whose teachers reported a similar constructivist approach to learning, in contrast to a direct instruction type of approach, showed better performance on word problem tests.

In a similar vein, there is much support for the value of drawing out student reasoning in the form of elaborated explanations as part of this conceptually oriented constructive learning process. For example, one of the best substantiated educational findings in cognitive science research is the educational benefit of explanation, and in particular, the self-explanation effect (Chi, 2000; Chi et al., 1994; Chi et al., 1989; Renkl, 2002). Self-explanation benefits learners by raising their awareness of their own knowledge gaps, abstracting problem specific knowledge into schemas that can be applied to other relevant cases, and elaborating the representation of knowledge in the learner's mind so that it can be more easily retrieved (VanLehn & Jones, 1993). The self-explanation effect appears to be related to the process of constructing an explanation. Webb (1985) observed that the cognitive benefits of explanations are restricted to elaborated explanations. When students possess sufficient background knowledge and are sufficiently motivated, presenting them with correctly worked example problems in math and science and directing them to "self-explain" has been extensively investigated and proven highly effective, even more effective than problem solving at early stages of skill acquisition (Atkinson, Renkl & Margaret, 2003; Renkl & Atkinson, 2003; Renkl, 2002; Schworm & Renkl, 2002; Lim & Moore, 2002; Renkl, Atkinson, Maier & Staley, 2002; Hummel & Nadolski, 2002; Bannert, 2002). Nevertheless, in typical classroom settings neither the appropriate level of background knowledge nor the ideal level of motivation can be assumed. Thus, an important question for improving the state of education is how to design interactions with instructional technology that are effective for supporting productive explanation activities in a way that would be practical to place in a realistic school setting.

Beyond the instructional value of explanation activities, student explanations and other verbal behavior have been demonstrated to be valuable for assessment. For example, student explanations have been a valuable source of insight into the evolution of student problem solving strategies in microgenetic studies of learning (Siegler, 1995; Siegler, 2002). Open-ended problems requiring deep reasoning and explanation go far beyond computational skills and provide deep insight into student understanding (Stylianou et al., 2000). The student's explanation makes the student's thinking visible to the teacher. An opinion poll of 180 teachers demonstrates that teachers highly value student explanations for assessment; Student explanations ranked third in terms of importance among the seven most frequently mentioned practices (Schmidt & Brosnan, 1996). However, assessment of student writing is time consuming, and teacher time is precious. The significant investment of time required to implement such activities in classrooms has often been prohibitive (Quinn & Wilson, 1997;

McIntosh & Draper, 2001). Thus, technology for automatic assessment of explanations, perhaps combining our content based approach with already established automatic essay grading technology (Foltz et al., 1998; Burstein et al., 1998), would enable a powerful tool for supporting constructivist instructional activities involving extended student explanations, which therefore require essay assessment, to be made more commonplace in schools. Furthermore, patterns observed in the recorded conversational interactions of students working in a collaborative setting were demonstrated to be highly predictive of student post-test performance (Webb, Nemer & Zuniga, 2002).

Nevertheless, we are not arguing that natural language by itself is the only important medium for assessment. For example, a recent study comparing assessments covering the same concepts using natural language, graphics, and equations showed that these different representations revealed different views of student knowledge, where sometimes misconceptions or knowledge gaps would appear in one representation and not the others (McCullough & Meltzer, 2001). Thus, our argument is that making use of multiple representations for the purpose of assessment including but not limited to natural language is valuable for gaining an accurate and complete picture of student understanding. However, in this article we focus specifically on assessment from natural language explanations.

Levels of Scaffolding

Earlier we described a continuum of sophistication in scientific reasoning and explanation. On one end of that continuum was understanding and explanation of isolated conceptual rules. At the other end was composition of multi-step lines of reasoning involving the selection and application of a sequence of conceptual rules in order to achieve a goal. As mentioned, our aim is to support students in constructing explanations as independently as possible. The dynamic nature of dialogue offers the flexibility of supporting students in a targeted way, using feedback to scaffold their process at their level of need. Here we make an argument for the empirical justification for adjustable levels of scaffolding tailored to the diagnosed level of student need on this continuum.

Although explanation activities have intrinsic cognitive benefits and students often learn from explanation activities without feedback, such as unsupported self-explanation during problem solving or worked example studying, many students provide explanations of low quality in these contexts (Renkl, Stark, Gruber, & Mandl, 1998). Feedback on explanation quality should support the generation of higher-quality explanations. Furthermore, in computer-based learning environments, such feedback is likely to provide a crucial incentive for students to take the explanation task seriously (Aleven & Koedinger, 2000). It is well-established that guidance and feedback help learning (Anderson, 1993, Ch. 7; Corbett & Anderson, 2001; Lee, 1992; Mathan & Koedinger, 2003; Schmidt & Bjork, 1992) and that unguided learning (e.g., pure discovery learning) is typically inefficient and ineffective (Mayer, 2004). Mathan and Koedinger (2003) specifically present empirical evidence in favor of an "intelligent novice" approach to scaffolding in which feedback is delayed in order to allow a controlled amount of unguided search on the student's part before help is offered.

Consistent with the results reported in (Mathan & Koedinger, 2003), we observed a similar pattern of benefit in connection with explanation activities. We observed a human tutor working

through the WHY2 curriculum, consisting of 10 qualitative physics problems such as The Pumpkin Problem. An analysis of the corpus of typed interactions between the students and the tutor in relation to knowledge gain as measured by a pre and post test provided support for encouraging students to proceed as independently as possible with their reasoning and explanation activities (Rosé et al., 2003b; Litman et al., 2004).

Our argument for scaffolding targeted at displayed level of need centers on an interpretation of a reliable correlation we found between average student turn length and learning with effect of pretest partialed out ($R=.515$, $p<.05$, $N=20$). In addition to the cognitive benefits of elaborated explanations discussed above, we observed that as students moved forward with their reasoning beyond the support offered by the scaffolding of tutor prompts, the more opportunities there were for students to make mistakes that yielded valuable opportunities for learning beyond those explicitly planned by the tutor.

We examined the relationship between student turn length and tutor feedback. There was a reliable correlation between average student turn length and average tutor turn length ($R=.585$, $p<.05$, $N=20$). We do not interpret this as an indication that students simply offered more explanation in response to longer tutor turns because of the distribution of tutor negative feedback. The likelihood of receiving some form of negative feedback in a tutoring turn strongly correlated with the length of the student's previous turn ($R=.8065$, $p<.01$)[2]. Thus, we conclude that elaborated student responses were rewarded by additional correction offered by the tutor, and that this additional feedback afforded by the student's venture into unprompted explanation contributed positively to the student's learning. This finding also explains an initially puzzling pattern we have observed on some pairs of pre/post tests where we found more misconceptions on the post test rather than less, as we had anticipated. Upon closer inspection we found that post test essays were far more elaborate than pretest essays. Thus, by supporting students in developing the ability to articulate their scientific reasoning, one outcome was that student misconceptions became more visible. Bringing misconceptions to the surface where they can be identified is one important step towards addressing them. Recent empirical results from the net based communication literature offer convincing evidence that providing experts with an inventory of a layperson's related conceptual knowledge enables them to construct explanations that are more readily understood and remembered (Wittwer, Nueckles, & Renkl, 2004).

For the tutor feedback analysis just discussed, we had coded tutor responses for types of negative feedback after an incorrect student answer. We explain that analysis here to show that our identification of instances of tutor negative feedback was reliable. We coded for three types of explicit negative feedback, which were not mutually exclusive. For illustrative purposes, we assume that the student has incorrectly claimed that gravitational force acts in the horizontal direction. Feedback marked as *Pointing* indicated that the tutor explicitly pointed out something wrong in the student's response, either by stating that it was wrong, directly questioning it, as in "Is gravity a horizontal force?", or by stating the opposite of what the student said, as in "Gravitational force does not act in the horizontal direction." *Negative* was assigned to a tutor

[2] To compute this correlation, we divided student turns from the complete transcripts of 7 students into 10 piles based on the length of the student's turn. We then computed the probability of receiving negative feedback by dividing the number of turns where the student received negative feedback from the tutor in the next turn over the total number of student turns. We then computed a simple linear regression between the median student turn length for each pile and the probability of receiving negative feedback.

response if the tutor began by saying, "No", or "That is not right." *Right* was assigned to a tutor response if the tutor corrected the student's incorrect statement in his turn, as in "Gravitational force acts in the *vertical* direction." *Ignore* was assigned to tutor responses to wrong answers that did not contain any of the above types of explicit negative feedback. These turns included cases where the tutor simply rephrased his original question or tried a different question altogether. We consider the *Ignore* responses to be implicit negative feedback. Using these detailed codings, we derived two coarser grained judgments, specifically negative feedback (i.e., any of the four) versus no negative feedback, and explicit (i.e., Pointing, Negative, or Right) versus implicit (i.e., Ignore).

We measured agreement on negative feedback coding over the same subset of the corpus used to evaluate the question type judgments. While, we did not achieve an acceptable Kappa measure over the most detailed judgments, both the negative feedback versus no negative feedback judgment (Kappa value of .78) and the explicit negative feedback versus implicit negative feedback judgment (Kappa value of .68) were reliable. Answers to 46% of tutor questions received some form of negative feedback. Negative feedback was most likely to be explicit (83% of the time).

Since our aim is to support students in constructing explanations of their scientific reasoning as independently as possible, we have developed technology to scaffold the process at different levels so that we can offer scaffolding at the level of the diagnosed need. In order to accomplish this, we make use of a technology that is easily adaptable to alternative levels of scaffolding. In particular, we utilize interactive tutorial dialogue technology that we refer to as Knowledge Construction Dialogues (KCDs) (Rosé et al., 2001a; Jordan, Rosé, & VanLehn, 2001). KCDs were motivated by the idea of Socratic tutoring. KCDs are interactive directed lines of reasoning that are each designed to lead students to learn as independently as possible one or a small number of concepts, thus implementing a preference for an "Ask, don't tell" strategy inspired by a previous empirical comparison between Socratic and Didactic style tutoring (Rosé, More, VanLehn, & Allbritton, 2001b). When a question is presented to a student, the student types a response in a text box in natural language. The student may also simply click on Continue, and thus neglect to answer the question. If the student enters a wrong or empty response, the system will engage the student in a remediation sub-dialogue designed to lead the student to the right answer to the corresponding question. The system selects a subdialogue based on the content of the student's response, so that incorrect responses that provide evidence of an underlying misconception can be handled differently than responses that simply show ignorance of correct concepts. Once the remediation is complete, the KCD returns to the next question in the directed line of reasoning. By manipulating the timing of when we offer tutorial dialogue support and the granularity and nature of the focus of the line of reasoning underlying the dialogue offered to students, we can use the same technology to offer scaffolding along the full spectrum of student needs.

In the original set of KCDs used in (Rosé et al., 2001a), we stepped students through lines of reasoning where we applied rules of physics in order to provide a foundation for students to understand and remember individual conceptual rules of physics. Students answered questions about things they experienced in their every day lives to help them understand. For example, "If you hold a book in your hand, which way do you feel the book pushing?" Students can answer these questions even if they don't know any physics. They simply require students to think about their experiences. And yet, these questions help them to see laws of physics at work. The

ultimate goal of these KCDs was to bring students to a place where they could remember and articulate a single rule of physics. Sometimes as part of that line of reasoning, students were asked to go one step further and apply those rules. For example, after discussing the concept of normal force applied by a horizontal object, students were asked to predict what would happen if the object was now tilted. If students were not able to make the conceptual leap, their understanding was scaffolded using a subdialogue, which is an embedded line of reasoning. Eventually, if the students were not able to apply the rule with help, the rule was applied for them. The focus of this work was to provide conceptual help when students displayed a gap in their understanding with faulty problem solving actions. Our evaluation showed that students who received this form of conceptual help were better able to answer questions about qualitative rules of physics on the post test (Rosé et a., 2001a).

The level at which tutor prompts in KCDs prompt behavior on the part of the student determines the level at which the KCD offers scaffolding to the student. In principle it is also possible to use KCDs to step students through multi-step chains of reasoning where they go beyond applying a rule and first perform the more difficult task of rule selection. For example, the tutor may ask a question such as "How would you determine the ball's displacement over a time interval if you know the ball's average speed?" Emphasizing this level of reasoning, we developed a class of KCDs called walk-through KCDs for this purpose. The primary purpose of these walk-through KCDs was to scaffold the process of selecting rules to apply in order to construct an appropriate line of reasoning for answering one of our conceptual physics problems. Nevertheless, since the ultimate goal is to bring students to a place where they can engage in multi-step scientific reasoning by themselves and communicate that reasoning in natural language, it was necessary to offer students the opportunity to do that and for the system to step in only when necessary. Encouraging students to engage in this process was the focus of the development of the WHY2 instructional process described below.

Instructional Process

In order to encourage students to engage in independent reasoning and explanation, when students begin to interact with WHY2, they are first presented with an essay question such as The Pumpkin Problem. After reading the essay question, the student types in an initial essay with no scaffolding. The system then analyzes the student's essay in order to detect the presence of misconceptions and required concepts, which are determined when each of WHY2's problems are designed. In this way, the system can assess the level of the student's need. The system then uses KCDs both for coaching students to insert missing required concepts (elicitation KCDs) and to remove the expression of misconceptions (remediation KCDs).

Consider the following example essay entered by a student in response to The Pumpkin Problem: "The pumpkin will land on the man. The pumpkin will have the same horizontal velocity as the man, thus an equal horizontal displacement at any time." This essay contains no evidence of the presence of misconceptions, however the reasoning is incomplete. For example, the student has not yet explicitly addressed the issue of which forces are acting on the pumpkin either while it is in the man's hands or after the release. Thus, the system engages the student in an elicitation KCD designed to bring this lack to the student's attention.

WHY2: You have the correct answer but what you have in your essay is not a sufficient explanation. I'll give you some hints about what is needed in the explanation. What forces are acting on the pumpkin after it is released?
Student: Gravity
WHY2: Yes. Since we know that gravity is the only force acting on the pumpkin, let me ask you this. As the gravitational force pulls the pumpkin down, does it also pull the pumpkin horizontally a little bit?
Student: No
WHY2: Fine. So is there any horizontal force on the pumpkin at all?
Student: Yes
WHY2: That's not right. Since gravity acts only in the vertical direction, then there is no force acting on the pumpkin in the horizontal direction. I believe this is an important point to cover in your essay. If you see what I mean, please fix your essay. Remember the submit button when you're finished

Elicitation KCDs, such as the one above, stand in contrast to remediation KCDs. They are not designed with the idea that the student has some wrong or even necessarily missing knowledge. All that can be confidently assumed from a missing piece of information in a student explanation is that the student did not recognize the need to include that piece of information. Nevertheless, if the student is in fact missing the associated piece of knowledge or has a related misconception, the student will not be able to respond appropriately to the elicitation KCD's prompts. We see an example of this in the student's final contribution to the elicitation KCD above. Once this knowledge gap is revealed, the system can remediate this faulty knowledge using a remediation subdialogue. In this case, a simple corrective explanation is offered in the tutor's final turn.

Once the elicitation KCD is complete, the student has the opportunity to revise the explanation. Here is the student's revised explanation: "The pumpkin will land on the man. The pumpkin will have the same initial horizontal velocity as the man. After being thrown, the only force acting on the pumpkin is gravity. The horizontal velocity of the man and the pumpkin remain the same." Note that the student has now addressed the issue regarding forces acting upon the pumpkin but has removed an important piece of argumentation related to the displacement of the pumpkin. An elicitation KCD is then used to address that missing knowledge.

WHY2: If two objects have the same horizontal velocity, and if they start out in the same position, then what will be the relationship between the horizontal displacement of the two objects from the point of release?
Student: same
WHY2: Good. So if the horizontal velocity of the pumpkin is the same as the horizontal velocity of the runner, then what is the relationship between their respective displacements from the point of release?
Student: They will equal each other

After this elicitation KCD, the student is then able to successfully add this information to the essay: "The pumpkin will land on the man. The pumpkin will have the same initial horizontal velocity as the man. After being thrown, the only force acting on the pumpkin is gravity. The

horizontal velocity of the man and the pumpkin remain the same. Both also started at the same point, and the horizontal displacement will be equal for both objects at all times." In this case the student did not demonstrate faulty knowledge about the connection between velocity and displacement. But if the student had, it could have been addressed with a remediation KCD designed to support a correct understanding of this concept. In this way elicitation KCDs can be used as a tool for identifying student misconceptions and missing knowledge. In the case of discovering such a lack, the system will engage the student in a remediation KCD to remediate the student's incorrect answer. Once the student has demonstrated the ability to articulate the desired concept, the elicitation KCD is complete, and the system asks the student to insert that required point in the essay. Here is an example remediation dialogue designed to help the student understand the connection between velocity and displacement.

> **WHY2:** Let's say a car is moving along a straight road with a constant velocity, Vc. Let's say the car moves along the x-axis. An airplane ascends on take-off with a velocity Va, at an angle, theta, to the x-axis. Say the car and airplane have the same (constant) velocity in the x-direction, or Vcx=Vax. For any given time interval, what is the relationship between the displacement of the car and the displacement of the airplane along the x-axis?
> **Student:** not sure
> **WHY2:** The displacement of a body in a given direction and over a given time interval is equal to what quantity multiplied by that time interval?
> **Student:** speed
> **WHY2:** The displacement of a body over a given time interval is equal to its average velocity multiplied by that time interval. If something moves at a constant velocity of magnitude Vc, what is the magnitude of its average velocity across any time interval?
> **Student:** Vc
> **WHY2:** Super. Thus, since the car and airplane move at the same constant velocity in the x-direction, what is the relationship between their average velocities in the x-direction over any interval of time?
> **Student:** same
> **WHY2:** Correct. So if the average velocity in the x-direction is the same for the car and airplane for any given time interval, what is the relationship between the displacements of the car and airplane in the x-direction?
> **Student:** same
> **WHY2:** Okay. In general, if two bodies have the same constant velocity along a given direction, for any time point, their displacements along that direction will be the same.

In general, remediation subdialogues are provided to students each time the student fails to give a correct answer to a tutor's question. Thus, main lines of reasoning have embedded remediation subdialogues, which in turn may have other embedded remediation subdialogues. This hierarchical structure allows KCDs to tailor their instruction to the specific needs of the student. After the student completes the remediation KCD by demonstrating a correct understanding of the underlying physics principle, the student is then asked to correct the essay where the

misconception was expressed. This revision cycle continues until no more flaws are found in the student's essay or some maximum number of revision cycles has been reached.

Motivation for Underlying Assessment Machinery

KCDs provide the scaffolding offered by WHY2, but before WHY2 can engage in the type of cyclic process described above, it must have an underlying assessment mechanism that allows it to select appropriate KCDs on each iteration. Our system is not the first to be faced with the task of assessing extended student explanations. For example, the well established area of automated essay grading has enjoyed a great deal of success at applying shallow language processing techniques to the problem of assigning general quality measures to student essays (Burstein et al., 1998; Foltz et al., 1998). Our task stands in contrast to this work, however, in that what we need to determine about student explanations is not simply a measure of quality, but we need a detailed assessment of where their specific demonstrated knowledge gaps are. The problem of providing reliable, detailed, content-based feedback to students is a more difficult problem than assigning a general quality measure. We are building the machinery to allow a fine grained level of detail in assessment of student explanations (Rosé & Hall, 2004; Makatchev, Jordan, & VanLehn, 2004). However, in this article we describe and evaluate the capabilities of a simplified mechanism that in the long run we plan to use mainly as a fall-back mechanism when more sophisticated processing fails.

Our simplified approach making use of text-classification technology builds upon other previous approaches that have addressed the problem of identifying individual pieces of content in extended explanations (Christie, 2003), sometimes called "answer aspects" (Wiemer-Hastings et al., 1998). In a subsequent section we argue why our hybrid approach is an advance over these previously published automatic assessment approaches.

Our text classification approach to scaffolding selection is a rough approximation to what we see our human tutor doing. Our ultimate goal is to provide the kind of tailored, focused feedback that we see our human tutoring offering students. Because elicitation KCDs can be used to probe for misconceptions, and since the high frequency of missing knowledge in student essays completely dwarfs the relatively infrequent occurrence of explicitly articulated misconceptions in our WHY2 corpus, as a simplifying measure we focus on the task of identifying correct expectations in the prescribed complete argumentation for each problem. As pointed out earlier, encouraging the development of the ability to articulate more complete lines of reasoning is one important step towards making student misconceptions visible and thus eventually to be identified and addressed by the system. Thus, for example, we search essays in response to The Pumpkin Problem for the following 5 expected reasoning steps, or Correct Answer Aspects (CAAs):

(1) After the release the only force acting on the pumpkin is the downward force of gravity.
(2) The pumpkin continues to have a constant horizontal velocity after it is released.
(3) The horizontal velocity of the pumpkin continues to be equal to the horizontal velocity of the man.

(4) The pumpkin and runner cover the same distance over the same time.
(5) The pumpkin will land on the runner.

A MATHEMATICAL FRAMEWORK FOR EVALUATION

In the previous section, we motivated our instructional approach, which is an iterative process of offering students the opportunity to develop or refine an explanation, identifying flaws in the student's work, selecting and presenting appropriate instructional interventions, and then offering students the opportunity to respond. The system's success in engaging in this type of interaction with the student depends upon it deriving an adequate image of the student's knowledge state from the student's verbal behavior. In this section we offer a mathematical model as a framework in which to evaluate the adequacy of the image provided by the technology for this purpose. The adequacy is measured in terms of the impact of the performance of that technology on the resulting appropriateness of the system's behavior. We argue that the model we provide applies to all cases of diagnosis under uncertainty, which in an intelligent tutoring context includes not only natural language interpretation but also complex and error prone mathematical or other formal language input, graphical input (i.e., diagrams, etc.), or gestures.

A student's verbal expression of reasoning offers a vivid picture of that student's knowledge state to the tutoring system in a way that would be understandable on multiple levels to a human tutor. Nevertheless, because even state-of-the-art language understanding technology cannot achieve a human tutor's sensitivity to the subtleties of the student's verbal behavior, the system's interpretation of the student's explanation can only be an approximation of the richness encoded there. The language understanding approach that is selected to do the assessment therefore provides the tutoring system with only a very grainy image of the student's knowledge state. The resolution of that image is determined both by the sophistication of the target representation and the accuracy of the encoding. The sophistication of the target representation is measured in terms of its representational power. In particular, more powerful representations are able to encode a wider range of distinctions and therefore may potentially select from a wider range of alternative views of the student. The higher the resolution of the image of the student's understanding, the greater the potential for the system to adapt to the student's needs. However, on the negative side if the accuracy of the encoding is low, greater representational power creates more ways in which the system's response to the student can be adapted in an inappropriate way. Thus, these two competing concerns must always be considered and balanced.

Rules or patterns that are matched against the target representation created by the selected language understanding approach can be compared to diagnostic medical tests used by doctors to test for the presence of disease. If rules always correspond to correct ideas, as in the approach presented in this paper, then the accuracy of the understanding approach can be computed by comparing the true set of Correct Answer Aspects (CAAs) to the identified set of CAAs. Such an evaluation is presented in Section 4. If a rule or pattern matches the created representation, the system has evidence that the associated idea was communicated by the student's language in the same way that medical diagnostic tests provide evidence but not proof of the presence of disease in a patient. Thus, we can make use of a similar mathematical model for evaluating the performance of a natural language interpretation approach to what is used in that community to

evaluate the effectiveness of diagnostic medical tests. Not coincidentally, a mathematical model originating with Signal Detection Theory typically used in that community for that purpose[3] has at its foundation similar measures to those that have been used in the text classification community to evaluate that work, namely, true positives, false positives, true negatives, and false negatives.

These four basic measures are used to define other measures that are more concretely relevant for the task. For example, Analysis Precision is the percentage of required points identified in the student essays that were actually present in those essays. In other words, it is the number of true positives divided by the sum of true positives and false positives. Related to this notion is Analysis False Alarm rate, which is the percentage of required points not present in the essay that were incorrectly identified by the system. In other words, this is the number of false negatives divided by the sum of true negatives and false negatives. Analysis Recall is the percentage of required points present in student essays that were actually identified by the system. This is the number of true positives divided by the sum of true positives and false negatives. We can use these same four values to define measures that are more closely related to what the student sees. For example, Intervention Selection Precision is the percentage of instructional interventions selected that are appropriate for the student. This is the number of true negatives divided by the sum of the true negatives and false negatives. A final measure not defined in terms of the four basic measures is essay completeness. This is the percentage of required points that are present in the student's essay.

In the medical field, the same four basic measures are used to define measures that are more directly relevant for their task, namely, Positive Predictive Value of a diagnostic test, Test Sensitivity, Specificity, and Prevalence of disease. By drawing an analogy between our measures and theirs, we can learn an important lesson for our type of diagnosis as well. In particular, the Positive Predictive Value of a diagnostic test depends both on what is called the Specificity as well as the Prevalence of the disease in the population. If disease is more prevalent, then the specificity of the test must be higher in order for the test to achieve an acceptable Positive Predictive Value. What is called Sensitivity in that field is analogous to our Analysis Recall. What they call Specificity is analogous to our Intervention Selection Precision. What they call Positive Predictive Value is analogous to our Analysis Precision. And their Negative predictive value is analogous to our Intervention Selection Precision. Prevalence of disease could be thought of as analogous to our student performance completeness. It is possible to verify that these pairs represent analogous quantities since both sets of measures are defined in terms of true positives, true negatives, false positives, and false negatives. What we find then is that as the completeness of student essays increases, the accuracy of our interpretation must be much higher in order to achieve an acceptable level of performance in terms of the selection of appropriate instructional interventions. Thus, the problem of high prevalence (or analogously, high student performance completeness) is a common problem across fields for tasks involving diagnosis under uncertainty.

Now we explore this problem in greater depth specifically with reference to intelligent tutoring. Broadly speaking, there are two ways in which the system's response can be inappropriately adapted to the student. Neglecting to give a student an instructional intervention that is needed means losing an opportunity to teach that student something that student needs to

[3] See for example http://www.epidemiolog.net/studymat/.

know. Conversely, offering an intervention that a student does not need means wasting a student's time, possibly distracting that student from what that student really needs to learn, and likely annoying or even confusing that student. Obviously the ideal is to build a system that maximizes the percentage of needed interventions that are appropriately provided while minimizing the percentage of unneeded interventions that are inappropriately provided. This maximizes the signal to noise ratio, and thus the overall relevance and completeness of the instruction offered to the student. Drawing again from Signal Detection Theory, to give an overall picture of the quality of the intervention selection enabled by an understanding approach, we will use d' as a measure. Intervention selection Recall and False Alarm Rate are used to compute d', which is the distance between two distributions, namely signal, and signal+noise. d' = Z(Recall) - Z(False Alarm Rate). Normally distributed Z values range between -2 and 2 about 95% of the time, so differences of twice that would be rare. Higher d' values are preferable to lower ones.

See Figure 1 for a summary of the measures used in our mathematical model. Note that A corresponds to false negatives, B corresponds to true positives, C corresponds to false positives, and D corresponds to true negatives. In terms of CAAs, note that A refers to CAAs emitted by the student but not detected by the tutor. B refers to CAAs emitted by the student that were correctly identified by the tutor. C refers to CAAs incorrectly identified by the tutor, since they were not emitted by the student. D refers to CAAs not emitted by the student and not identified by the tutor.

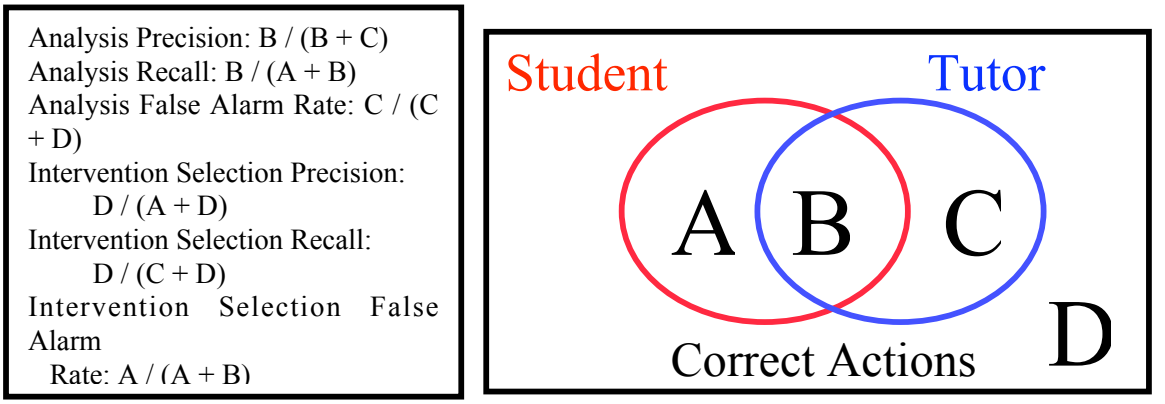

Fig.1. Mathematical Model of the relationship between analysis accuracy and intervention selection accuracy.

Taking a close look at the equations in Figure 1, you will notice that intervention selection accuracy is a function of analysis accuracy and student performance completeness. Figure 2 illustrates how drastically intervention selection accuracy drops as student performance completeness increases even with an analysis precision and recall of 90%.

It is also possible to show that, holding intervention selection accuracy constant, that analysis accuracy varies with student performance completeness. But in practice, it makes more

sense to consider analysis accuracy as inherent in the selected analysis approach, and intervention selection accuracy is derived from that. Note that we define student performance completeness here as the percentage correct student actions that the student has performed, whether they are recognized or not. It does not take into consideration any expressions of misconceptions or wrong actions also present.

Now let's consider precisely how intervention selection precision, recall, and false alarm rate are related to analysis precision, recall, and false alarm rate as well as student performance completeness in terms of CAAs. As you see from the equations in Figure 1, intervention selection precision is the number of interventions correctly given divided by the total number of interventions given. Interventions are given whenever a correct student action is not identified. Thus, when analysis recall is low, a lot of correct student actions that the student has performed will not be identified. Thus, the corresponding interventions will be incorrectly given. A needed intervention is not given whenever a student action is incorrectly identified. Thus, when analysis precision is low, many of the actions that are identified have not actually been performed by the student. Thus, many of the interventions that the student needs will not be given.

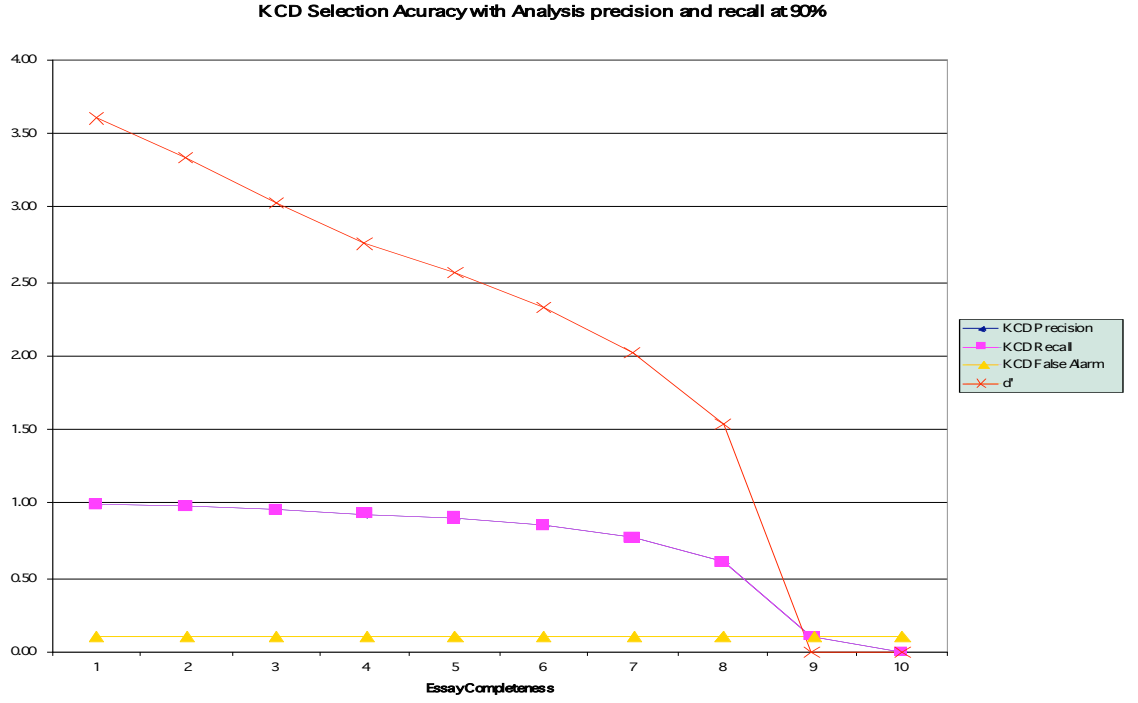

Fig.2. This graph illustrates how intervention selection precision, recall, and d' vary with student performance completeness, keeping 0.90 analysis precision and 0.90 analysis recall. Note that intervention selection precision and recall are indistinguishable in this graph.

When student performance completeness is high, D will be low. If student performance completeness is 70%, D can be no greater than 30%. Remember that intervention selection precision is D / (A + D). If analysis recall is perfect, then A will be 0%, so intervention selection precision will be 100%. But if analysis recall is less than perfect, even a little bit less, then intervention selection precision goes down very fast. Since D is small, even if A is equally small, intervention selection precision is down to 50%. Intervention selection recall is determined by the relationship between C, the points not in the essay that were identified, and D, the points not in the essay that were not identified, in particular, D / (C + D). Thus, if analysis precision is perfect, then C will be 0%, so intervention selection recall will be 100%. But if it is not perfect, then it will bring the intervention selection recall down fast, again because D is necessarily small.

The opposite is the case when student performance completeness is very low, lets say 20%. In this case, A and B must both be low. Thus, even if analysis recall is 0%, A will be no more than 20%. Let's say that analysis recall and precision are around 86%. Whenever student performance completeness is low and analysis precision is reasonably high, A and C must both be small in comparison with D. Analysis precision and recall of 86% would have been disastrous for a high performance completeness student as we saw above. However, in this case, it would mean that A is 3%, B is 17%, C is 3%, and D is 77%. Thus, intervention selection precision and recall are both 96%. Therefore, the situation is quite different when student performance completeness is low. If analysis accuracy is reasonably high, intervention selection accuracy will also be high.

In summary, when student performance completeness is high, C and D are very small. As a result, intervention selection precision D/(A+D) and recall D/(C+D) are more sensitive to analysis accuracy as essay completeness increases. Therefore, intervention selection precision and recall will tend to decrease as student performance completeness increases. As intervention selection recall decreases, so does the difference between Recall and False Alarm Rate. The final result is that the overall quality of intervention selection, as computed by d', decreases as student performance completeness increases. In particular, even if Analysis Precision and Recall are almost perfect, specifically at 90%, then Intervention Selection Precision, Recall, and False Alarm Rate become unsatisfyingly low once essay completeness is 70% or higher. See Figure 2. Notice also how d' drops quickly as student performance completeness increases.

Thus, when we depend upon our analysis of the student's performance to tell us how to select interventions for students, we may be at a loss with respect to helping high end students. One might wonder whether it is necessary to worry about the high end students since it is often said that good students will learn even with a bad teacher. Nevertheless, no matter how good the students are, if we fail to present them with instruction on the topics that they lack, they will not have the opportunity to learn those topics. Additionally, we have evidence in the form of expressions of frustration in our WHY2 log files that students are unhappy when they receive help on many topics that they do not need help on. However, we do not yet have a definitive answer on what is the minimum acceptable level of intervention selection accuracy. This is still under investigation.

STUDENT ESSAY ANALYSIS AS A TEXT CLASSIFICATION PROBLEM

Earlier we motivated the use of text classification technology for student essay assessment and provided a mathematical model for evaluating its adequacy. We argued based on this mathematical model why it is so important that the competing concerns of representational power and encoding accuracy must always be carefully balanced when selecting a language understanding approach to use for essay assessment. In this section we describe and evaluate our specific text classification approach in light of this model. One contribution of this work is the demonstration of the robustness of a novel hybrid text classification approach called CarmelTC. As typical of text classification approaches, CarmelTC encodes each text as a vector of feature values. It then learns rules, based on labelled examples, which classify novel texts based on learned patterns of feature values. The CarmelTC text classification approach combines the typical use of words as features with the unique use of features extracted from a deep syntactic functional analysis. Our evaluation demonstrates the contribution of these features to the overall robustness of the approach. We argue why specific characteristics of the conceptual physics domain and other domains with a similar character (i.e., causal domains) create a need for these additional features. Thus, another contribution of this work is the demonstration of the importance of not treating text classification technology as a black box. Specifically, we argue that careful selection of features for classification based on domain considerations is useful for optimizing text classification performance for student explanation assessment.

Overview of the Approach

CarmelTC is a text classification approach, and as with any other text classification approach, one must decide what to use as classes. We took the very coarse grained approach of classifying each sentence as expressing one of the correct expectations associated with a problem, or nothing if the sentence does not correctly and completely express any of the associated expectations. The correct expectations for each of our WHY2 problems were formulated in such a way that they are normally expressed in student essays in a single sentence. Earlier we justified the selection of 5 specific required points, or Correct Answer Aspects (CAAs), for a correct and complete student response to The Pumpkin Problem. Based on these analyses, we allow for a total of six alternative classifications for each text segment:

> **[Class 1]** Sentence expresses the idea that after the release the only force acting on the pumpkin is the downward force of gravity.
> **[Class 2]** Sentence expresses the idea that the pumpkin continues to have a constant horizontal velocity after it is released.
> **[Class 3]** Sentence expresses the idea that the horizontal velocity of the pumpkin continues to be equal to the horizontal velocity of the man.
> **[Class 4]** Sentence expresses the idea that the pumpkin and runner cover the same distance over the same time.
> **[Class 5]** Sentence expresses the idea that the pumpkin will land on the runner.
> **[Class 6]** Sentence does not adequately express any of the above specified key points.

Distinguishing between these 6 classes is a challenging text classification problem. Notice that with these six classes, in some cases what distinguishes sentences from one class and sentences from another class is very subtle. For example, "Thus, the pumpkin's horizontal velocity, which is equal to that of the man when he released it, will remain constant." belongs to Class 2 although it could easily be mistaken for Class 3. Similarly, "So long as no other horizontal force acts upon the pumpkin while it is in the air, this velocity will stay the same.", belongs to Class 2 although looks similar on the surface to either Class 1 or 3. A related problem is that sentences that should be classified as "nothing" may look very similar on the surface to sentences belonging to one or more of the other classes. For example, "It will land on the ground where the runner threw it up." contains all of the words required to correctly express the idea corresponding to Class 5, although it does not express this idea, and in fact expresses a wrong idea. "Bag of Words" approaches have trouble making these subtle distinctions between sentences. They make their decision based on which words are present in a text, and in these cases, the words to express sentences in one class are very close or even identical to those typically used to express sentences in another class. This in addition to the fact that "bag of words" approaches work best on large portions of text, much larger even than the whole essays that students working with Why2 type, puts "bag of words" approaches at a disadvantage for applications like this one.

Note that this classification task is strikingly different from those typically used for evaluating text classification systems. First, these classifications represent specific whole propositions rather than general topics, such as those used for classifying web pages (Craven et al., 1998), namely "student", "faculty", "staff", etc. Secondly, the texts are much shorter, i.e., one sentence in comparison with a whole web page, which is a disadvantage for "bag of words" approaches.

Combining Deep and Shallow Approaches to Understanding

CarmelTC, the text classification approach described in this paper, is a hybrid approach involving both symbolic and "bag of words" techniques. The idea of developing a hybrid language understanding approach is not original with CarmelTC. It is part of the logical progression of experimentation with alternative approaches for language understanding in dialogue based tutoring systems. Traditionally, many successful tutoring systems that accept natural language input have employed shallow approaches to language understanding. For example, CIRCSIM-tutor (Glass, 1999) and Andes-Atlas (Rosé et al, 2001a) parse student answers using shallow semantic grammars to identify key concepts embedded therein. These systems have restricted their scope to primarily handling responses to short answer questions, which can be handled well with these shallow analysis approaches. The Auto-Tutor (Wiemer-Hastings et al, 1998) system uses Latent Semantic Analysis (LSA) to process lengthy student answers. "Bag of Words" approaches such as LSA (Landauer et al., 1998) HAL (Burgess et al., 1998), and Rainbow (McCallum, 1996), have enjoyed a great deal of success in a wide range of applications. However, they gloss over key aspects of meaning that are communicated structurally through scope and subordination and ignore common, but nevertheless crucial, function words such as 'not'. It has been demonstrated to perform poorly in highly causal domains, such as Experimental Design (Malatesta et al., 2002).

Because of the limitations of these shallow language understanding approaches, recently a number of dialogue based tutoring systems have begun to employ more linguistically sophisticated techniques for analyzing student language input, namely the Geometry tutor (Aleven et al., 2003), and BEETLE (Core & Moore, 2004; Zinn et al., 2002) in addition to WHY2. Nevertheless, both the shallow and the deep approaches to language understanding have their own unique strengths and weaknesses. "Bag of Words" approaches require relatively little development time, are totally impervious to ungrammatical input, and tend to perform well because much can be inferred about student knowledge just from the words they use. On the other hand, symbolic, knowledge based approaches require a great deal of development time and tend to be more brittle than superficial "Bag of Words" types of approaches, although robustness techniques can increase their level of imperviousness (Rosé & Lavie, 2001). To their credit, linguistic knowledge based approaches are more precise and capture nuances that "Bag of Words" approaches miss.

Recent work suggests that symbolic and "Bag of Words" approaches can be productively combined. For example, syntactic information can be used to modify the LSA space of a verb in order to make LSA sensitive to different word senses (Kintsch, 2001). Along similar lines, syntactic information can be used, as in Structured Latent Semantic Analysis (SLSA), to improve the results obtained by LSA over single sentences (Wiemer-Hastings & Zipitria, 2001).

The WHY2 CarmelTC approach was initially developed as a fallback approach for the more sophisticated symbolic language understanding approach involving both deep syntactic and semantic sentence level interpretation (Rosé, 2000) as well as discourse level interpretation (Makatchev, Jordan, & VanLehn, 2004). The original architecture of the WHY2 symbolic language understanding approach is described in (Vanlehn et al., 2002). In brief, it uses the CARMEL core understanding component for symbolic sentence level language understanding (Rosé, 2000). It takes natural language as input and produces a set of first order logical forms to pass on to the discourse language understanding (DLU) module. After sentence level processing, the DLU module combines the sentence level information by making abductive inferences about how the pieces of information fit together using Tacitus-Lite+ (Jordan et al., 2003). The resulting proof trees are then used as the basis for determining which required points are missing from student essays, when optional points are not mentioned or inferable from what is mentioned, and which misconceptions may be present.

What is unique about CarmelTC is that it combines features from a deep syntactic analysis provided by the CARMEL core understanding component (Rosé, 2000) with predictions from the Rainbow Naive Bayes text classifier (McCallum, 1996). The CARMEL core understanding approach provides the facilities to develop domain specific semantic knowledge sources that allow it to construct a semantic interpretation for each input sentence. However, in CarmelTC we make use only of its deep syntactic analysis of each input sentence. From this syntactic analysis we extract individual features that encode functional relationships between syntactic heads (e.g., (subj-throw man)), tense information (e.g., (tense-throw past)), and information about passivization and negation (e.g., (negation-throw +) or (passive-throw -)). We also extract word features that indicate the presence or absence of a root form of a word from the sentence. Rainbow has been used for a wide range of text classification tasks. With rainbow, P(doc,Class), i.e., the probability of a document belonging to class Class, is estimated by multiplying P(Class), i.e., the prior probability of the class, by the product over all of the words w_i found in the text of P(w_i|Class), i.e., the probability of the word given that class. This product is normalized over

the prior probability of all words. Using the individual features extracted from the deep syntactic analysis of the input as well as the "bag of words" Naïve Bayes classification of the input sentence, CarmelTC builds a vector representation of each input sentence, with each vector position corresponding to one of these features. We then use the ID3 decision tree learning algorithm (Mitchell, 1997) to induce rules for identifying sentence classes based on these feature vectors.

With this hybrid approach, our goal has been to keep as many of the advantages of both the symbolic approach and the "bag of words" classification approach as possible while avoiding some of the pitfalls of each. Since the CarmelTC approach does not use the syntactic analysis as a whole, it does not require that the system be able to construct a totally complete and correct syntactic analysis of the student's text input. It can very effectively make use of partial parses. Thus, it is more robust than the purely symbolic approach. And since it makes use only of the syntactic analysis of a sentence, rather than also making use of a semantic interpretation, it does not require any sort of domain specific knowledge engineering. The "bag of words" classification of the sentence gives one opinion, if you will, of how the sentence should be classified. One can think of CarmelTC as learning rules for deciding whether it should believe this interpretation or select a different one.

CarmelTC is most similar to the text classification approach described in (Furnkranz et al., 1998). In the approach described in (Furnkranz et al., 1998), features that note the presence or absence of a word from a text as well as extraction patterns from AutoSlog-TS (Riloff, 1996) form the feature set that are input to the Ripper (Cohen, 1995), which learns rules for classifying texts based on these features. While CarmelTC is similar in spirit both in terms of the sorts of features used and the general sort of learning approach, CarmelTC is different from (Furnkranz et al., 1998) in several respects. Where (Furnkranz et al., 1998) make use of AutoSlog-TS extraction patterns, CarmelTC makes use of features extracted from a deep syntactic analysis of the text. Thus, the syntactic features extracted from CARMEL are more general. For example, for the sentence "The force was applied by the man to the object", CARMEL assigns the same functional roles as for "The man applied the force to the object" and also for the noun phrase "the man that applied the force to the object". Since AutoSlog-TS performs a surface syntactic analysis, it would assign a different representation to all aspects of these texts where there is variation in the surface syntax. Note that while computing a deep syntactic analysis is more computationally expensive than computing a surface syntactic analysis, we can do so very efficiently using an incrementalized version of LCFlex that takes advantage of student typing time (Rosé et al., 2002). Like (Furnkranz et al., 1998), we also extract word features that indicate the presence or absence of a root form of a word from the text. Additionally, for CarmelTC one of the features for each training text that is made available to the rule learning algorithm is the classification obtained using the Rainbow Naïve Bayes classifier (McCallum, 1996). Because the texts classified with CarmelTC are so much shorter than those of (Furnkranz et al., 1998), the feature set provided to the learning algorithm was small enough that it was not necessary to use a learning algorithm as sophisticated as Ripper (Cohen, 1995). Thus, we used ID3 (Mitchell, 1997) instead with excellent results.

Using Features From CARMEL's Deep Syntactic Analysis

The symbolic features used for the CarmelTC approach are extracted from a deep syntactic functional analysis constructed using the CARMEL grammar (Rosé, 2000) and the broad coverage COMLEX lexicon (Grishman et al., 1994), containing 40,000 lexical items. For parsing we use an incrementalized version of the LCFlex robust parser (Rosé et al., 2002), which was designed for efficient, robust interpretation.

Syntactic feature structures produced by the grammar factor out those aspects of syntax that modify the surface realization of a sentence but do not change its deep functional analysis. These aspects include tense, negation, mood, modality, and syntactic transformations such as passivization and extraction. In order to do this reliably, the component of the grammar that performs the deep syntactic analysis of verb argument functional relationships was generated automatically from a feature representation for each of COMLEX's verb subcategorization tags. It was verified that the 91 verb subcategorization tags documented in the COMLEX manual were covered by the encodings, and thus by the resulting grammar rules (Rosé et al., 2002). These tags cover a wide range of patterns of syntactic control and predication relationships. Each tag corresponds to one or more case frames. Each case frame corresponds to a number of different surface realizations due to passivization, relative clause extraction, and wh-movement. Altogether there are 519 configurations of a verb in relation to its arguments that are covered by the 91 subcategorization tags, all of which are covered by the grammar.

There are nine syntactic functional roles assigned by the grammar. These roles include subj (subject), causesubj (causative subject), obj (object), iobj (indirect object), pred (descriptive predicate, like an adjectival phrase or an adverb phrase), comp (a clausal complement), modifier, and possessor. The roles pertaining to the relationship between a verb and its arguments are assigned based on the subcat tags associated with verbs in COMLEX. However, in some cases, arguments that COMLEX assigns the role of subj (subject) get redefined as causesubj (causative subject). For example, the subject in "the pumpkin moved" is just a subject but in "the man moved the pumpkin", the subject would get the role causesubj instead since 'move' is a causative-inchoative verb and the obj (object) role is filled in the second case. Aspectual verbs, such as "start" and "continue", are also treated differently. They take a verb phrase as their object, but in the deep syntactic analysis, they are treated as temporal modifiers of their objects. And their subjects are treated as the subjects not of the aspectual verb but of the verb phrase that is their object. The modifier role is used to specify the relationship between any syntactic head and its adjunct modifiers. Possessor is used to describe the relationship between a head noun and its genitive specifier, as in "man" in either "the man's pumpkin" or "the pumpkin of the man". Note that with eventive nouns such as "toss" in "the man's toss of the pumpkin", "man" would get the role of subj (subject) instead, and "pumpkin" would get the role of obj (object). Light verbs, such as "to have" and "to be", are treated compositionally by the grammar, but they can be interpreted paraphrastically by adding construct entries describing specific idiomatic uses of them such as "to have a clue" or "to be equal to" (Rosé, 2000).

The complete set of features extracted from all training example texts are used to build a vector representation of each text. ID3 refers to this vector in order to select those features that

are most informative for classifying example texts into one of the available classes or another. Thus, it learns rules of the same form listed as examples in Figure 3. Note that these rules are slight simplifications on the actual rules learned. These rules illustrate how CarmelTC is able to overcome some of the difficulties that "bag of words" approaches face in making the sorts of fine grained distinctions that are necessary for the Why2 text classification task. For example, the sentence "It will land on the ground where the runner threw it up", which should be classified as "nothing", i.e., Class 6, includes all of the same words as "It will land on the runner", which is correctly classified as Class 5. In this case, though, the syntactic features extracted from the analysis of this sentence indicate that the modifier of "where" is "throw". Thus, the landing is where the pumpkin was thrown and not where the runner is after the throw. Thus, even this one syntactic feature is enough to inform CarmelTC that the assigned class of 5 is not correct and that Class 6 is the correct class instead.

Table 1
Example CarmelTC Rules

Text	**Actual Class**	**Class assigned based on words**	**CarmelTC Rule**
It will land on the ground where the runner threw it up.	Class6	Class5	(bow class5) + (mod-where throw) --> Class6
So long as no other horizontal force acts upon the pumpkin while it is in the air, this velocity will stay the same.	Class2	Class1 or Class3	(bow class1) + (marker-act so-long-as) + (possesor-velocity nil) --> Class2

EVALUATION

Initial Evaluation

We conducted an evaluation to compare the effectiveness of CarmelTC at analyzing student essays in comparison to LSA, Rainbow, and a purely symbolic approach similar to (Furnkranz et al., 1998), which we refer to here as CarmelTCsymb. CarmelTCsymb is identical to CarmelTC except that it does not include in its feature set the prediction from Rainbow. Thus, by comparing CarmelTC with Rainbow and LSA, we can demonstrate the superiority of our hybrid approach to purely "bag of words" approaches. And by comparing with CarmelTCsymb, we can demonstrate the superiority of our hybrid approach to an otherwise equivalent purely symbolic approach. Thus, we observe that the features extracted from the parse of the sentence and the prediction based on the words only each contribute to the overall accuracy of the CarmelTC approach.

We conducted our evaluation over a corpus of 126 previously unseen student essays in response to the Pumpkin Problem described above, with a total of 500 text segments, and just under 6000 words altogether. We first tested to see if the text segments could be reliably tagged by humans with the six possible Classes associated with the problem. Note that this includes

"nothing" as a class, i.e., Class 6. Three human coders hand classified text segments for 20 essays. We computed a pairwise Kappa coefficient to measure the agreement between coders, which was always greater than .7, thus demonstrating acceptable agreement. We then selected two coders to individually classify the remaining sentences in the corpus. They then met to come to a consensus on the tagging. The resulting consensus tagged corpus was used as a gold standard for this evaluation. Using this gold standard, we conducted a comparison of the four approaches on the problem of tallying the set of "correct answer aspects" present in each student essay.

The LSA space used for this evaluation was trained over three first year physics text books. The other three approaches are trained over a corpus of tagged examples using a 50 fold random sampling evaluation, similar to a cross-validation methodology. On each iteration, we randomly selected a subset of essays such that the number of text segments included in the test set were greater than 10 but less than 15. The randomly selected essays were then used as a test set for that iteration, and the remainder of the essays were used for training in addition to a corpus of 248 hand tagged example sentences extracted from a corpus of human-human tutoring transcripts in our domain. The training of the three approaches differed only in terms of how the training data was partitioned.

Rainbow and CarmelTCsymb were trained using all of the example sentences in the corpus as a single training set. CarmelTC, on the other hand, required partitioning the training data into two subsets, one for training the Rainbow model used for generating the value of its Rainbow feature, and one subset for training the decision trees. This is because for CarmelTC, the data for training Rainbow must be separate from that used to train the decision trees so the decision trees are trained from a realistic distribution of assigned Rainbow classes based on its performance on unseen data rather than on Rainbow's training data.

In setting up our evaluation, we made it our goal to present our competing approaches in the best possible light in order to provide CarmelTC with the strongest competitors as possible. Note that LSA works by using its trained LSA space to construct a vector representation for any text based on the set of words included therein. It can thus be used for text classification by comparing the vector obtained for a set of exemplar texts for each class with that obtained from the text to be classified. We tested LSA using as exemplars the same set of examples used as Rainbow training data, but it always performed better when using a small set of hand picked exemplars. Thus, we present results here using only those hand picked exemplars.

For every approach except LSA, we first segmented the essays at sentence boundaries and classified each sentence separately. However, for LSA, rather than classify each segment separately, we compared the LSA vector for the entire essay to the exemplars for each class (other than "nothing"), since LSA's performance is better with longer texts. We verified that LSA also performed better specifically on our task under these circumstances. Thus, we compared each essay to each exemplar, and we counted LSA as identifying the corresponding "correct answer aspect" if the cosine value obtained by comparing the two vectors was above a threshold. We tested LSA with threshold values between .1 and .9 at increments of .1 as well as testing a threshold of .53 as is used in the AUTO-TUTOR system (Wiemer-Hastings et al., 1998). As expected, as the threshold increases from .1 to .9, recall and false alarm rate both decrease together as precision increases. We determined based on computing f-scores for each threshold level that .53 achieves the best trade off between precision and recall. Thus, we used a threshold of .53, to determine whether LSA identified the corresponding key point in the student essay or not for the evaluation presented here.

We evaluated the four approaches in terms of precision, recall, false alarm rate, and f-score, which were computed for each approach for each test essay, and then averaged over the whole set of test essays. We computed precision by dividing the number of "correct answer aspects" (CAAs) correctly identified by the total number of CAAs identified. For essays containing no CAAS, we counted precision as 1 where none were identified and 0 otherwise. We computed recall by dividing the number of CAAs correctly identified over the number of CAAs actually present in the essay. For essays containing no CAAs, we counted recall as 1 for all approaches. False alarm rate was computed by dividing the number of CAAs incorrectly identified by the total number of CAAs that could potentially be incorrectly identified. For essays containing all possible CAAs, we counted false alarm rate as 0 for all approaches. F-scores were computed using 1 as the beta value in order to treat precision and recall as equally important.

The results presented in Table 2 demonstrate that CarmelTC outperforms the other approaches. In particular, CarmelTC achieves the highest f-score, which combines the precision and recall scores into a single measure. In comparison with CarmelTCsymb, CarmelTC achieves a higher recall as well as a slightly higher precision. While LSA achieves a slightly higher precision, its recall is much lower. Thus, the difference between the two approaches is shown most clearly in the f-score value, which favors CarmelTC. Rainbow achieves a lower score than CarmelTC in terms of precision, recall, false alarm rate, and f-score.

Table 2
This Table compares the performance of the 4 alternative approaches in the per essay evaluation in terms of precision, recall, false alarm rate, and f-score

	Precision	**Recall**	**False Alarm Rate**	**F-score**
LSA	93%	54%	3%	.70
Rainbow	81%	73%	9%	.77
CarmelTCSymb	88%	72%	7%	.79
CarmelTC	90%	80%	8%	.85

A More Challenging Evaluation

After our initial evaluation of CarmelTC on the identification of problem specific expectations, we tested it on a larger data set of 1324 sentences extracted from essays for 5 different qualitative problems, and a larger, more abstract set of 15 classes that represent problem independent physics rules, such as:

> **[Class1]:** If A and B move together, they will have the same acceleration, velocity and displacement at all times.
> **[Class2]:** Two objects with the same velocity have the same displacements at all times.
> **[Class3]:** If an object has non-zero velocity, it has non-zero displacement.

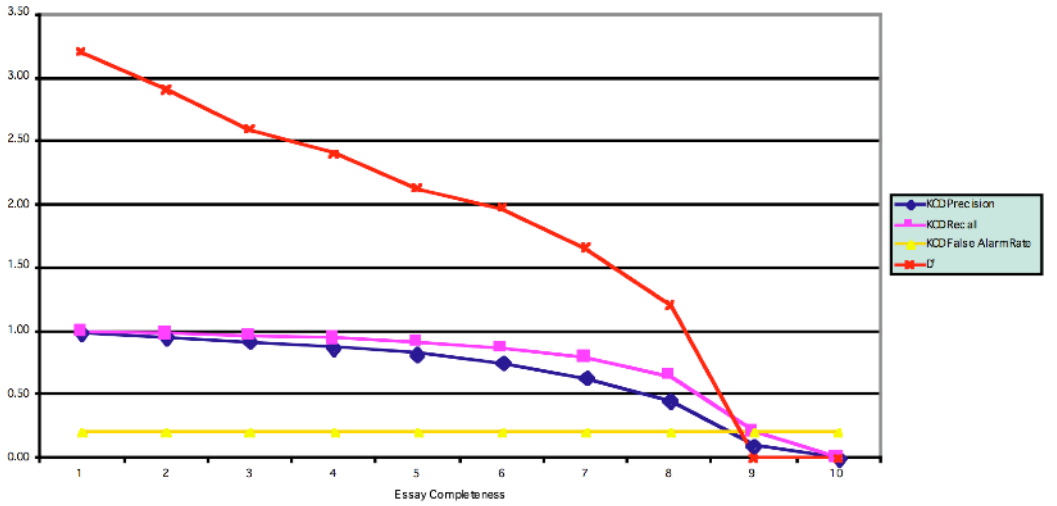

Fig.3. This graph illustrates how intervention selection precision, recall, and d' vary with student performance completeness, using our CarmelTC results.

Notice that these classes are defined at a more abstract, problem independent level than those used in the previous evaluation. In particular, they do not refer to the exact objects used in specific problems. Furthermore, certain pieces of information encoded in these rules can be encoded in language in very different ways depending upon the problem. For example, here is an example of Class1 from the Pumpkin problem:

"Since the runner's speed is constant, he and the pumpkin always have the same horizontal displacement, and the pumpkin will land in your hand."

Now observe an example of the same class from a problem about a package that was released from a moving airplane (The Airplane Problem):

"If the package is flying with the airplane, then it is traveling with it and the moment it dropped it still has that initial velocity given to it by the airplane."

Notice that these two sentences have very different ways of communicating the idea that two objects are traveling together at the same velocity. Note also that the contribution of features above the word level does not help to bridge the gap in this case.

The classification problem was made more complex by assigning classes to sentences taking into consideration information implicit in the discourse context but not made explicit in the sentence itself. For example, look at this sentence from the Airplane Problem of Class 3:

"Hence the packet does not hit the target."

And in some cases, part of the rule is implicit in the text. For example, in the following example, the idea of the airplane having a nonzero velocity is known only because an airplane must be moving in order to stay in the air:

"The packet will not hit the target if it is dropped when the airplane is right above the pre-selected target."

We trained the Rainbow models on 1324 tagged sentences. We then evaluated its performance on a completely separate set of 1324 sentences. It's error rate over the test set was 49%. We then evaluated CarmelTC over the evaluation set using a 10 fold cross evaluation, and the error rate was only 37%, which is a 25% reduction in error rate. While this error rate is higher than that achieved in the previous, less challenging evaluation, it is still substantially better than the bag-of-words Rainbow naïve bayes classifier. Thus, this evaluation provides further evidence of the importance of deep syntactic functional roles in the interpretation of sentences in highly causal domains.

CONCLUSIONS AND CURRENT DIRECTIONS

In this paper we have explored the problem of selecting appropriate interventions for students based on an analysis of their performance. In particular, we have introduced the WHY2 Qualitative Physics intelligent tutoring system that presents short essay questions to students and coaches them to improve their qualitative physics explanations in response to these questions. In this context, we present CarmelTC, a hybrid text classification approach for analyzing student essays and thus selecting needed interventions for students. We have described our KCD technology, in particular describing the elicitation and remediation KCDs that we provide to students as interventions in order to coach them to improve their essays as well as our evidence of the effectiveness of this technology in teaching qualitative physics concepts to students. We have also described a mathematical model that demonstrates a general problem with selecting interventions based on an analysis of student performance in circumstances where there is uncertainty with this interpretation, such as with natural language input, complex and error prone mathematical or other formal language input, graphical input (i.e., diagrams, etc.), or gestures. In particular, when student performance completeness is high, intervention selection precision and recall are more sensitive to analysis accuracy, and increasingly so as performance completeness increases. Therefore, intervention selection accuracy will tend to decrease as student performance completeness increases. Our mathematical model allows us to make precise and explicit exactly the extent to which this problem occurs. In light of this model, we have evaluated CarmelTC over data collected from students interacting with our system. Our evaluation demonstrates that the novel hybrid CarmelTC approach outperforms both LSA and Rainbow, as well as a purely symbolic approach.

ACKNOWLEDGEMENTS

This research was supported by the Office of Naval Research, Cognitive and Neural Sciences Division under grant number N00014-0-1-0600 and by NSF grant number 9720359 to CIRCLE, Center for Interdisciplinary Research on Constructivist Learning Environments and the University of Pittsburgh and Carnegie Mellon University.

REFERENCES

Aleven, V., & Koedinger, K. R. (2002). An Effective Meta-cognitive Strategy: Learning by Doing and Explaining with a Computer-Based Cognitive Tutor. *Cognitive Science,* 26(2), 147-179.

Aleven, V., & Koedinger, K. R. (2000). The Need for Tutorial Dialog to Support Self-Explanation. In C. P. Rosé & R. Freedman (Eds.) Building Dialogue Systems for Tutorial Applications, Papers of the 2000 AAAI Fall Symposium (pp. 65-73). Technical Report FS-00-01. Menlo Park, CA: AAAI Press.

Aleven V., Koedinger, K. R., & Popescu, O. (2003). A Tutorial Dialogue System to Support Self-Explanation: Evaluation and Open Questions. In U. Hoppe, F. Verdejo & J. Kay (Eds.) *Proceedings of the 11th International Conference on Artificial Intelligence in Education, AI-ED 2003* (pp. 39-46). Amsterdam: IOS.

Aleven, V., Koedinger, K., & Cross, K. (1999). Tutoring answer-explanation fosters learning and understanding. In S. P. Lajoie & M. Vivet (Eds.) *Proceedings of the 9th International Conference on Artificial Intelligence in Education* (pp. 199-206). Amsterdam: IOS.

Aleven, V., Koedinger, K., Sinclair, H. C., & Snyder, J. (1998). Combating shallow learning in a tutor for geometry problem solving. In B. P. Goettl, H. M. Halff, C. L. Redfield & V. J. Shute (Eds.) *Intelligent Tutoring Systems, the Fourth International Conference,* (pp. 364-373). Berlin: Springer Verlag.

Anderson, J. R. (1993). *Rules of the Mind.* Hillsdale, NJ: Lawrence Erlbaum.

Ashley, K. D., Desai, R., & Levine, J. M. (2002). Teaching Case-Based Argumentation Concepts Using Dialectic Arguments vs. Didactic Explanations. In S. A. Cerri, G. Gouardères & F. Paraguaçu (Eds.) *Proceedings of Sixth International Conference on Intelligent Tutoring Systems, ITS 2002* (pp. 585-595). Berlin: Springer Verlag.

Atkinson, R., Renkl, A., & Margaret, M. (2003). Transitioning From Studying Examples to Solving Problems: Effects of Self-Explanation Prompts and Fading Worked-Out Steps. *Journal of Educational Psychology*, 95(4), 774-783.

Bannert, M. (2002). Managing cognitive load – recent trends in cognitive theory. *Learning & Instruction Special Issue: Cognitive Load Theory,* 12(1), 139-146.

Baumert, J., & Köller, O. (2000). Unterrichtsgestaltung, verständnisvolles Lernen und multiple Zielerreichung im Mathematik- und Physikunterricht der gymnasialen Oberstufe. In J. Baumert, W. Bos & R. Lehmann (Eds.) *TIMSS/III: Dritte Internationale Mathematik- und Naturwissenschaftsstudie - Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Bd. 2 - Mathematische und physikalische Kompetenzen in der Oberstufe* [Classroom instruction, insightful learning, and multiple educational objectives in mathematics and physics classes in the upper academic track. In J. Baumert, W. Bos & R. Lehmann (Eds.) TIMSS/III: The Third International Mathematics and Science Study – Mathematical and scientific literacy at the end of schooling: Volume 2 – Mathematical and physics competencies in the upper academic track (pp. 271-315). Opladen: Leske + Budrich.

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

Bork, A. (2004). Distance Learning Today and Tomorrow. In G. Kearsley (Ed.) *Online Learning: Reflections on the Transformation of Education* (pp. 26-41). Englewood Cliffs, NJ: Educational Technology Publications.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in Context Space: Words, Sentences, Discourse. *Discourse Processes*, 25(2), 211-257.

Burstein, J., Kukich, K., Wolff, S., Chi, L., & Chodorow, M. (1998). Enriching automated essay scoring using discourse marking. *Proceedings of the Workshop on Discourse Relations and Discourse Marking* (pp. 206-210). Association of Computational Linguistics.

Carbonell, J. (1970). AI in CAI: an artificial intelligence approach to computer-assisted instruction. *IEEE Transactions on Man-Machine Systems,* 11(4), 190-202.

Carbonell, J. (1969). On man-computer interaction: a model and some related issues. *IEEE Transactions on Systems Science and Cybernetics,* 5(1), 16-26.

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.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3), 439-477.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2), 145-182.

Chi, M., de Leeuw, N., Chiu, M., & La Vanchar, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18 (3), 439-477.

Christie, J. R. (2003). Automated essay marking for content: Does it work? In *Proceedings of the CAA Conference.*

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational Outcomes of Tutoring: A meta-analysis of Findings. *Americal Educational Research Journal,* 19*,* 237-248.

Cohen, W. (1995). Fast effective rule induction. *Proceedings of the 12th International Conference on Machine Learning* (pp. 115-123). ACM Press.

Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: impact on learning rate, achievement, and attitudes. *Proceedings of CHI 2001* (pp. 245-252). New York: ACM Press.

Core, M., & Moore, J. D. (2004). Robustness versus Fidelity in Natural Language Understanding. *Proceedings of the Second International Workshop on Scalable Natural Language Understanding* (pp. 1-8). Association for Computational Linguistics.

Craven, M., DiPasquio, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., & Slattery, S. (1998). Learning to extract symbolic knowledge from the World Wide Web. *Proceedings of the 15th National Conference on Artificial Intelligence* (pp. 509-516). AAAI Press.

Duit, R., & Confrey, J. (1996). Reorganizing the curriculum and teaching to improve learning in science and mathematics. In D. F. Treagust, R. Duit & B. J. Fraser (Eds.) *Improve teaching and learning in science and mathematics* (pp. 79-93). New York: Teachers College Press.

Evens, M., & Michael, J. (2004). *One-on-One Tutoring by Humans and Machines*. Lawrence Erlbaum and Associates.

Foltz, P., Kintsch, W., & Landauer, T. (1998). The Measurement of Textual Coherence with Latent Semantic Analysis. *Discourse Processes*, 25, 285-308.

Freedman, R. (2000). Using a Reactive Planner as the Basis for a Dialogue Agent. *Proceedings of FLAIRS 2000*, Orlando, 203-208, AAAI Press.

Furnkranz, J., Mitchell, T., Mitchell, M., & Riloff, E. (1998). A Case Study in Using Linguistic Phrases for Text Categorization on the WWW. *Proceedings from the AAAI/ICML Workshop on Learning for Text Categorization* (pp. 5-12). AAAI Press.

Gertner, A., & VanLehn, K. (2000). Andes: A Coached Problem Solving Environment for Physics. In G. Gauthier, C. Frasson & K. VanLehn (Eds) *Intelligent Tutoring Systems: 5th International Conference* (pp. 133-142). Lecture Notes in Computer Science, Vol. 1839. Berlin: Springer.

Glass, M. (1999). *Broadening Input Understanding in an Intelligent Tutoring System*. PhD thesis, Illinois Institute of Technology.

Graesser, A., VanLehn, K., the TRG, & the NLT (2002). *Why2 Report: Evaluation of Why/Atlas, Why/AutoTutor, and Accomplished Human Tutors on Learning Gains for Qualitative Physics Problems and Explanations*. LRDC Tech Report, University of Pittsburgh.

Graesser, A., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., and the Tutoring Research Group, (2000). Using Latent Semantic Analysis to Evaluate the Contributions of Students in AutoTutor. In a special issue of *Interactive Learning Environments*, J. Psotka, guest editor, 8, 129-148.

Graesser, A. C., Bowers, C. A., Hacker, D.J., & Person, N. K. (1998). An anatomy of naturalistic tutoring. In K. Hogan & M. Pressley (Eds.) *Scaffolding of instruction.* Brookline Books.

Grishman, R., Macleod, C., & Meyers, A. (1994). COMLEX syntax: Building a computational lexicon. *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)* (pp. 268-272). Association for Computational Linguistics.

Hake, R. (1998). Interactive engagement versus traditional methods: A six-thousand student survey of mechanics test data for introductory physics students. *American Journal of Physics,* 66, 64-74.

Halloun, I., & Hestenes, D. (1985). The initial knowledge state of college physics students, *American Journal of Physics*, 53(11), 1043-1055.

Heffernan, N. T., & Koedinger, K. R. (2002). An intelligent tutoring system incorporating a model of an experienced human tutor. In S. A. Cerri, G. Gouardères & F. Paraguaçu (Eds.) *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems, ITS 2002* (pp. 596-607). Berlin: Springer Verlag.

Hewett, P. (1987). *Conceptual Physics for Everyone,* Addison-Wesley Publishing Company.

Hummel, H., & Nadolski, R. (2002). Cueing for schema construction: Designing problem-solving multimedia practicals, *Contemporary Educational Psychology,* 27(2), 229-249.

Jordan, P., Makatchev, M., & VanLehn, K., (2003). Abductive Theorem Proving for Analyzing Student Explanations. In U. Hoppe, F. Verdejo & J. Kay (Eds.) *Proceedings of the 11th International Conference on Artificial Intelligence in Education*, *AI-ED 2003* (pp. 73-80). Amsterdam: IOS Press.

Jordan, P., Rosé, C. P., & VanLehn, K. (2001). Tools for Authoring Tutorial Dialogue Knowledge. In J. D. Moore, C. L. Redfield & W. L. Johnson (Eds.) *Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future, Proceedings of AI-ED 2001* (pp. 222-233). Amsterdam: IOS Press.

Kintsch, W. (2001). Predication. *Cognitive Science,* 25, 173-202.

Koedinger, K. R., Anderson, J.R., Hadley, W.H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education,* 8, 30-43.

Kunter, M., & Baumert, J. (2004). Constructivist Approaches in the Secondary School Mathematics Classroom and Their Effects on Students' Learning, Interest, and Sense of Challenge: a Re-Analysis of the German TIMSS Data. *Proceedings of the 1st IEA International Research Conference*, Lefkosia, Cyprus.

Landauer, T. K., Foltz, P. W., & Latham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Lee, A. Y. (1992). Using tutoring systems to study learning. *Behavior Research Methods, Instruments, and Computers,* 24(2), 205-212.

Lim, E., & Moore, D. (2002). Problem solving in geometry: Comparing the effects of non-goal specific instruction and conventional worked examples. *Educational Psychology,* 22(5), 591-612.

Litman, D., Bhembe, D., Rosé, C., Forbes-Riley, K., Silliman, S. & VanLehn, K. (2004). Spoken Versus Typed Human and Computer Dialogue Tutoring. In *Proceedings of the Intelligent Tutoring Systems Conference* (pp. 368-379). Berlin: Springer Verlag.

Malatesta, K., Wiemer-Hastings, P., & Robertson, J. (2002). Beyond the Short Answer Question with Research Methods Tutor. In S. A. Cerri, G. Gouarderes & F. Paraguacu (Eds.) *Intelligent Tutoring Systems, 2002: 6th International Conference* (pp. 562-573). Berlin: Springer Verlag.

Makatchev, M., Jordan, P., & VanLehn, K. (2004). Abductive Theorem Proving for Analyzing Student Explanations and Guiding Feedback in Intelligent Tutoring Systems. *Journal of Automated Reasoning for Special Issue on Automated Reasoning and Theorem Proving in Education*, 32(3), 187-226.

Maloney, D. (1994). Research on problem solving: Physics. In D. L. Gabel (Ed.) *Handbook of Research on Science Teaching and Learning* (pp. 327-354). Macmillan: New York.

Mathan, S., & Koedinger, K. R. (2003). Recasting the Feedback Debate: Benefits of Tutoring Error Detection and Correction Skills. In U. Hoppe, F. Verdejo & J. Kay (Eds.) *Proceedings of the 11th International Conference on Artificial Intelligence in Education*, *AI-ED 2003* (pp. 13-203). Amsterdam: IOS Press.

McCallum, A. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/~mccallum/bow.

McIntosh, M. E., & Draper, R. J. (2001). Using learning logs in mathematics: writing to learn. *Mathematics Teacher,* 94(7), 554-557.

McCullough, L., & Meltzer, D. (2001). Differences in male/female response patterns on alternative-format versions of FCI items. *Proceedings of the Physics Education Research Conference* (pp. 103-106). Rochester, New York, July 25-26.

Meltzer, D., & Manivannan, K. (2002). Transforming the lecture hall environment: The Fully Interactive Physics Lecture. *American Journal of Physics*, 72, 1432-1446.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

Quinn, R. J., & Wilson, M. (1997). Writing in the Mathematics Classroom: Teacher Beliefs and Practices. *Clearing House,* 71, 14-20.

Renkl, A., & Atkinson, R. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational Psychologist,* 38(1), 15-22.

Renkl, A. (2002). Learning from worked-out examples: Instructional explanations supplement self-explanations. *Learning & Instruction,* 12, 529-556.

Renkl, A., Atkinson, R., Maier, U., & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. *Journal of Experimental Education*, 70(4), 293-315.

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.

Riloff, E. (1996). Automatically generating extraction patterns from untagged text. *Proceedings of the 13th National Conference on Artificial Intelligence* (pp.1044-1049). AAAI Press.

Rosé, C. P., & Hall, B. (2004). A Little Goes a Long Way: Quick Authoring of Semantic Knowledge Sources for Interpretation. *Proceedings of the 2nd International Workshop on SCalable Natural Language Understanding, SCaNaLu '04,*(pp. 17-24). Association for Computational Linguistics.

Rosé, C. P., Bhembe, D., Siler, S., Srivastava, R., & VanLehn, K. (2003b). The Role of Why Questions in Effective Human Tutoring, In U. Hoppe, F. Verdejo & J. Kay (Eds.) *Proceedings of the 11th International Conference on Artificial Intelligence in Education, AI-ED 2003* (pp. 55-62). Amsterdam: IOS Press.

Rosé, C. P., Roque, A., Bhembe, D., & VanLehn, K. (2002). An Efficient Incremental Architecture for Robust Interpretation. *Proceedings of the Human Languages Technologies Conference* (pp. 307-311). San Diego, California, Morgan Kaufman Publishers.

Rosé, C. P., Jordan, P., Ringenberg, M., Siler, S., VanLehn, K., & Weinstein, A. (2001a). Interactive Conceptual Tutoring in Atlas-Andes. In J. D. Moore, C. L. Redfield & W. L. Johnson (Eds.) *Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future, Proceedings of AI-ED 2001* (pp. 256-266). Amsterdam: IOS Press.

Rosé, C. P., Moore, J. D., VanLehn, K., & Allbritton, D. (2001b). A Comparative Evaluation of Socratic versus Didactic Tutoring. *Proceedings of Cognitive Sciences Society* (pp. 869-874).

Rosé, C. P., & Lavie, A. (2001). Balancing robustness and efficiency in unification-augmented context-free parsers for large practical applications. In J. C. Junqua & G. Van Noord (Eds.) *Robustness in Language and Speech Technology* (pp. 240-266). Amsterdam: Kluwer.

Rosé, C. P. (2000). A Framework for Robust Semantic Interpretation, *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 311-318). Association for Computational Linguistics.

Schmidt, M. E., & Brosnan, P. A. (1996). Mathematics Assessment: Practices and Reporting Methods. *School Science and Mathematics*, 96(1), 17-20.

Schworm, S., & Renkl, A. (2002). Learning by solved example problems: Instructional explanations reduce self-explanation activity. In W. D. Gray & C. D. Schunn (Eds.) *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 816-821). Mahwah, NJ: Erlbaum.

Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. *Cognitive Psychology*, 28, 225-273.

Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31-58). New York: Cambridge University.

Staub, F. C., & Stern, E. (2002). The nature of teachers' pedagogical content beliefs matters for students' achievement gains: Quasi-experimental evidence from elementary mathematics. *Journal of Educational Psychology*, 94(2), 344-355.

Stevens, A., & Collins, A. (1977). The Goal Structure of a Socratic Tutor. In *Proceedings of the National ACM Conference*. Association for Computing Machinery, New York, (Also available as *BBN Report No. 3518* from Bolt Beranek and Newman Inc., Cambridge, Mass., 02138).

Stylianou, D. A., Kenny, P. A., Siler, E. A., & Alacaci, C. (2000). Gaining insight into students' thinking through assessment tasks. *Mathematics Teaching in the Middle School,* 6(2), 136-144.

VanLehn, K., Jordan, P., Rosé, C. P., and The Natural Language Tutoring Group, (2002). The Architecture of Why2-Atlas: a coach for qualitative physics essay writing. In S. A. Cerri, G. Gouarderes & F. Paraguacu (Eds.) *Intelligent Tutoring Systems, 2002: 6th International Conference* (pp. 158-167). Berlin: Springer.

VanLehn, K., Freedman, R., Jordan, P., Murray, C., Rosé, C. P., Schulze, K., Shelby, R., Treacy, D., Weinstein, A. & Wintersgill, M. (2000). Fading and deepening: The next steps for Andes and other model-tracing tutors. In G. Gauthier, C. Frasson & K. VanLehn (Eds) *Intelligent Tutoring Systems: 5th International Conference* (pp. 474-483). Lecture Notes in Computer Science, Vol. 1839. Berlin: Springer.

VanLehn, K., & Jones, R. M. (1993). Better learners use analogical problem solving sparingly. In R. S. Michalski & G. Tecuci (Eds.) Proceedings of the Second International Workshop on Multistrategy Learning (pp. 19-30). Fairfax, VA: Center for Artificial Intelligence. Also in P. E. Utgoff (Ed.) *Proceedings of the Tenth International Conference on Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Webb, N., Nemer, K., & Zuniga, S. (2002). Short Circuits of Superconductors? Effects of Group Composition on High-Achieving Students' Science Assessment Performance, *American Educational Research Journal*, 39(4), 943-989.

Webb, N. (1985). *Learning to cooperate, cooperating to learn*. New York: Plenum Publishing.

Wiemer-Hastings, P. (2000). Adding syntactic information to LSA. *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society* (pp. 989-993). Mahwah, NJ: Lawrence Erlbaum Associates.

Wiemer-Hastings, P., Graesser, A., Harter, D. and the Tutoring Research Group (1998). The Foundations and Architecture of AutoTutor. In B. Goettl, H. Halff, C. Redfield & V. Shute (Eds.) *Intelligent Tutoring Systems: 4th International Conference (ITS '98)* (pp334-343). Berlin: Springer-Verlag.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In S. P. Lajoie & M. Vivet (Eds.) *Proceedings of the 9th International Conference on Artificial Intelligence in Education* (pp. 535-542). Amsterdam: IOS Press.

Wiemer-Hastings, P., & Zipitria, I. (2001). Rules for syntax, vectors for semantics. *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society*, 1112-1117.

Wittwer, J., Nückles, M., Renkl, A. (2004). Can experts benefit from information about a layperson's knowledge for giving adaptive explanations? In K. Forbus, D. Gentner & T. Regier (Eds.) *Proceedings Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp.1464-1469). Mahwah, NJ: Lawrence Erlbaum Associates.

Zinn, C., Moore, J. D., & Core, M. G. (2002). A 3-Tier Planning Architecture for Managing Tutorial Dialogue. In S. A. Cerri, G. Gouardères & F. Paraguaçu (Eds.) *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems, ITS 2002* (pp. 574-584). Berlin: Springer Verlag.