

A Hybrid Language Understanding Approach for Robust Selection of Tutoring Goals

Carolyn P. Rosé, Dumisizwe Bhembe, Antonio Roque, Stephanie Siler,
Ramesh Srivastava, and Kurt VanLehn

LRDC, University of Pittsburgh
Pittsburgh, PA 15260 USA
rosecp@pitt.edu

Abstract. In this paper we explore the problem of selecting appropriate Knowledge Construction Dialogues (KCDs) for the purpose of encouraging students to include important points in their qualitative physics explanations that are missing. We describe a hybrid symbolic/statistical approach developed in the context of the WHY2 conceptual physics tutor (Vanlehn et al., 2002). Our preliminary results demonstrate that our hybrid approach outperforms both the symbolic approach and the statistical approach by themselves.

1 Introduction

Recent studies of human tutoring suggest that a productive activity for teaching is to have students explain physical systems qualitatively (Chi et al, 1981). The goal of the WHY2 project (Vanlehn et al., 2002) is to coach students as they explain physics systems in natural language in response to short essay questions such as, “Suppose you are running in a straight line at constant speed. You throw a pumpkin straight up. Where will it land? Explain.” The WHY2 system has at its disposal a library of knowledge construction dialogues (KCDs), i.e., interactive directed lines of reasoning, each of which is designed either to elicit a specific idea (i.e., an elicitation KCD) or to remediate a specific misconception (i.e., a remediation KCD).

In order to do an effective job at selecting appropriate KCDs in response to student essays, the system must first analyze the student’s essay to assess which key points are missing from the student’s argument and which misconceptions the student may have based on what the student said. Thus, in this paper we explore a hybrid symbolic/statistical approach for robustly diagnosing student essay answers to qualitative physics questions in the context of the WHY2 conceptual physics tutor. In order to build an effective system, it is important both to give students the KCDs that they do need, and to avoid giving them extraneous KCDs that they do not need. Our preliminary results reported in this paper demonstrate that in both respects our hybrid approach outperforms both the symbolic approach and the statistical approach by themselves.

2 The Role of KCDs in Why2

When students interact with WHY2, they are first presented with an essay question. After reading the essay question, the student types in an initial essay. 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. 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).

Fig. 1. This figure contains an example remediation KCD for teaching the connection between velocity and displacement.

1-Tutor Let's say a car is moving along a straight road with a constant velocity, V_c . Let's say the car moves along the x-axis. An airplane ascends on take-off with a velocity V_a , at an angle, θ , to the x-axis. Say the car and airplane have the same (constant) velocity in the x-direction, or $V_{c_x} = V_{a_x}$.

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?

2-Student not sure

3-Tutor 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?

4-Student speed

5-Tutor The displacement of a body over a given time interval is equal to its average velocity divided by that time interval.

If something moves at a constant velocity of magnitude V_c , what is the magnitude of its average velocity across any time interval?

6-Student V_c

7-Tutor 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?

8-Student same

9-Tutor 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?

10-Student same

11-Tutor 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.

The ultimate goal of WHY2 is to coach students through the process of constructing explanations that are complete and do not contain any misconceptions. Knowledge construction dialogues (KCDs) are the main instructional tools used by WHY2 for coaching students. KCDs are interactive directed lines of reasoning that each teach a specific concept. Through these KCDs, students are lead

step by step by a sequence of questions designed to guide them into discovering conceptual knowledge as independently as possible. They have a very simple underlying dialogue management mechanism, specifically a finite state push down automaton. And they make use of very simple shallow semantic parsing grammars to analyze student input, classifying it into one of a small number of pre-defined answer classes. KCDs have already been demonstrated to be effective for increasing students' conceptual knowledge of physics (Rosé et al., 2001).

To illustrate the revision cycle that WHY2 engages students in, take as an example the case where the student has the misconception that two objects with the same velocity and initial position may have different displacements from their points of origin. This can be remediated by engaging them in a dialogue about the relationship between velocity and displacement, as in Figure 1. Typical of KCDs, this KCD begins by introducing a scenario to make the underlying physics concept easy for students to relate to. The tutor's turn ends with a question. In Turn 2, the student indicates that he does not know the answer. Turns 3-10 form a subdialogue meant to respond to the lack of knowledge communicated by the student's Turn 2. In general, remediation subdialogues are provided to students each time the student fails to give a correct answer to a tutor's question. Thus, with main lines of reasoning having embedded remediation subdialogues, which in turn may have other embedded remediation subdialogues, KCDs follow a hierarchical structure. 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.

Fig. 2. This Figure contains an example Elicitation KCD about the connection between velocity and displacement.

- 1-Tutor** 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?
- 2-Student** same
- 3-Tutor** 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?
- 4-Student** they will equal each other

If the system determines that a required concept is absent from a student's essay, the system engages the student in an elicitation KCD in order to encourage the student to articulate that concept. An example elicitation KCD is found in Figure 2. In this case the system is attempting to encourage the student to include a statement about the displacements of two objects being equal because their respective velocities are equal. Elicitation KCDs are typically shorter than

Fig. 3. This Figure summarizes our model for predicting KCD precision, recall, and false alarm rate from analysis precision, recall, and false alarm rate.

Recall:	$\text{NumberCorrectlyIdentified} / \text{CorrectPoints}$
Precision:	$\text{NumberCorrectlyIdentified} / \text{NumberIdentified}$
IncorrectPoints	$\text{TotalNumberPoints} - \text{CorrectPoints}$
IncorrectlyIdentified	$\text{NumberIdentified} - \text{NumberCorrectlyIdentified}$
False alarm rate:	$\text{IncorrectlyIdentified} / \text{IncorrectPoints}$
EssayQuality:	$\text{CorrectPoints} / \text{TotalNumberPoints}$
NumberCorrectlyIdentified:	$\text{Recall} * \text{CorrectPoints}$
NumberIdentified:	$\text{NumberCorrectlyIdentified} / \text{Precision}$
TotalKcdsGiven:	$\text{TotalNumberPoints} - \text{NumberIdentified}$
KcdsNeeded:	$\text{TotalNumberPoints} - \text{CorrectPoints}$
CorrectButNotIdentified:	$\text{CorrectPoints} - \text{NumberCorrectlyIdentified}$
KcdsCorrectlyGiven:	$\text{TotalKcdsGiven} - \text{CorrectButNotIdentified}$
KCD recall:	$\text{KcdsCorrectlyGiven} / \text{KcdsNeeded}$
KCD precision:	$\text{KcdsCorrectlyGiven} / \text{TotalKcdsGiven}$
KCDsNotNeeded	$\text{TotalNumberPoints} - \text{KcdsNeeded}$
KCDsIncorrectlyGiven	$\text{TotalKcdsGiven} - \text{KcdsCorrectlyGiven}$
KCD false alarm rate:	$\text{KCDsIncorrectlyGiven} / \text{KCDsNotNeeded}$

remediation KCDs. The idea behind them is that the student may already know the idea that they are meant to elicit and just has neglected to mention it in the essay. So they are short and ask questions meant to prompt the student to articulate the desired concept. If the student does not in fact know the desired concept, then the student will not be able to answer the questions correctly. 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.

3 Selecting Appropriate KCDs

In order to build an effective system, it is important both to give students the KCDs that they do need, and to avoid giving them extraneous KCDs that they do not need. Neglecting to give a student a KCD that is needed means losing an opportunity to teach that student something that student needs to know. Giving a KCD 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. Thus, we would like to build a system with a high KCD recall and low KCD false alarm rate, where we define KCD recall as the percentage of KCDs that a student needs that the system gives. And KCD false alarm rate as the percentage of KCDs that the student does not need that the system gives.

Nevertheless, analyzing student essays is a computational linguistics problem, and performance on this task is most naturally measured in terms of analysis precision, recall, and false alarm rate over a corpus of student essays. Analysis precision is the percentage of required points and misconceptions identified in the student essays that were actually present in those essays. Note that this is undefined in the case that no required points are identified. 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. Analysis recall is the percentage of misconceptions and required points present in student essays that were actually identified by the system. Note that this is undefined whenever there are no required points present in a student essay. Naturally, a system that is good at accurately identifying required points and misconceptions in student essays will also be good at selecting appropriate KCDs to engage students in. However, the relationship between analysis precision, recall, and false alarm rate and KCD precision, recall, and false alarm rate varies widely depending upon the quality of student essays. Thus, in order to make valid predictions about student experience with the system based on experiments over corpora of previously collected student essays, we built a mathematical model to compute KCD precision, recall, and false alarm rate from analysis precision, recall, and false alarm rate as it varies with different essay qualities. The model is summarized in Figure 3. From this model it is possible to predict how well we need to do at analyzing student essays in order to do a good job at selecting appropriate KCDs. It also makes it possible to make informed decisions about which out of a set of alternative language understanding approaches is most suitable based on their relative levels of analysis precision, recall, and false alarm rate.

We define Recall for analysis as the number of required points that WHY2 correctly identifies as present in a student essay (`NumberCorrectlyIdentified`) divided by the total number of required points actually present in the essay (`CorrectPoints`). Precision is `NumberCorrectlyIdentified` divided by the total number of required points that WHY2 identified, correctly or incorrectly (`Num-`

Fig. 4. This Table illustrates how KCD precision and recall vary with essay quality, keeping 0.90 analysis precision and 0.90 analysis recall.

Essay Quality	KCD Precision	KCD Recall
0.10-0.30	0.98	0.98
0.40-0.70	0.86	0.86
0.80-1.00	0.23	0.23

berIdentified). The number of points not correctly encoded in an essay (IncorrectPoints) is computed by subtracting CorrectPoints from TotalNumberPoints. To compute the false alarm rate, then, simply subtract NumberCorrectlyIdentified from NumberIdentified and divide the resulting number by IncorrectPoints. EssayQuality is CorrectPoints divided by the total number of required points (TotalNumberPoints).

In order to project KCD precision, recall, and false alarm rate for different essay qualities, we need to transform these equations in order to compute values for NumberCorrectlyIdentified and NumberIdentified as they vary with essay quality. Thus, from the Recall equation we derive the equation that NumberCorrectlyIdentified is Recall multiplied by CorrectPoints. And from the Precision equation we derive the equation that NumberIdentified equals NumberCorrectlyIdentified divided by Precision. WHY2 gives an elicitation KCD for every required point not identified in the student essay. Thus, the total number of elicitation KCDs given correctly or incorrectly (TotalKcdsGiven) is TotalNumberPoints minus NumberIdentified. However, the number of KCDs that the student actually needs (KcdsNeeded) is TotalNumberPoints minus CorrectPoints. In order to determine how many KCDs were correctly given (KcdsCorrectlyGiven), we first need to know how many required points the student included in the essay that were not identified by WHY2 (CorrectButNotIdentified). If we know CorrectPoints and NumberCorrectlyIdentified, we can get CorrectButNotIdentified by subtracting NumberCorrectlyIdentified from CorrectPoints. Then, KcdsCorrectlyGiven will be TotalKcdsGiven - CorrectButNotIdentified, since a KCD will be incorrectly given if the student expressed the corresponding point but WHY2 missed it. Now we have enough information to compute KCD precision and recall. KCD recall is the percentage of KCDs that were given that the student needed, thus, KcdsCorrectlyGiven divided KcdsNeeded. And KCD precision is the percentage of KCDs given that were actually needed, thus, KcdsCorrectlyGiven divided by TotalKcdsGiven. To compute KCD false alarm rate, you must first determine the number of KCDs not needed (KCDsNotNeeded). You can compute this by subtracting KcdsNeeded from TotalNumberPoints. You also need to know how many KCDs were incorrectly given (KCDsIncorrectlyGiven). You can compute this by subtracting KcdsCorrectlyGiven from TotalKcdsGiven. Note that this is equivalent to CorrectButNotIdentified. Thus, KCD false alarm rate is KCDsIncorrectlyGiven divided by KCDsNotNeeded. Note that the pro-

Fig. 5. KCD precision and recall with 0.88 analysis Precision, 0.75 analysis Recall, and .08 analysis False Alarm Rate. Note that this is the result we get with our best combined approach to essay analysis described below in the Results section.

Essay Quality	KCD Precision	KCD Recall
0.10-0.30	0.94	0.97
0.40-0.70	0.72	0.86
0.80-1.00	0.13	0.22

jection of analysis precision and recall onto KCD precision and recall works out most accurately if we treat the undefined cases for analysis precision and recall discussed above as 1.0.

From this mathematical model we determined that as essay quality increases, it becomes much more difficult to do a good job at selecting appropriate KCDs for students. In fact, selecting appropriate KCDs for students with essay qualities of 0.80 or higher may well be completely out of our reach. In particular, even if analysis precision and recall are at 0.90, KCD precision, recall, and false alarm rate become unsatisfyingly low once essay quality is 0.70 or higher. See Figure 4. Thus, helping excellent students improve their ability to construct high quality conceptual physics explanations may require an entirely different approach. From this model we have also determined, not too surprisingly, that performance can remain reasonable even if analysis recall is low. A low analysis precision means students will not get KCDs that are needed. On the other hand, a low recall means that students will get KCDs that they do not need. In Figure 5 we see that if analysis recall is low but precision remains near the 0.90 level, KCD recall remains high. Although this phenomenon seems counter-intuitive at first glance, it makes sense when one considers that if precision remains the same but recall is decreased, then the total number of points identified will be smaller, thus the total number of KCDs given will be higher. When essay quality is low and many KCDs are needed, the likelihood is that increasing the number of KCDs given will increase the number of KCDs correctly given. Nevertheless, KCD precision seriously suffers for higher quality essays. By the time essay quality is at 0.40, a quarter of the KCDs given will be inappropriate, and over half of the KCDs given for essays of quality 0.70 or more will be inappropriate.

4 Combining Deep and Shallow Approaches to Language Understanding

Many successful tutoring systems that accept natural language input employ shallow approaches to language understanding. For example, CIRCUSIM-TUTOR (Glass, 1999) and Andes-Atlas (Rosé et al., 2001) parse student answers using shallow semantic grammars to identify key concepts embedded therein. 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. 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 (Alevan et al., 2001), BEETLE (Core et al., 2001), and WHY2 (Vanlehn et al., 2002). Each approach has its 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 2000). To their credit, linguistic knowledge based approaches are more precise and capture nuances that “Bag of Words” approaches miss. For example, they capture key aspects of meaning that are communicated structurally through scope and subordination and do not ignore common, but nevertheless crucial, function words such as ‘not’.

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, 2002). 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 and Zipitria, 2001).

A detailed description of our approach to language understanding is beyond the scope of this paper, but can be found in (Vanlehn et al., 2002). In brief, we use the CARMEL core understanding component (Rosé, 2000) for symbolic sentence level language understanding. 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 (Jordan et al., 2002). We use Rainbow (McCallum, 1996), a naive Bayes classifier, for an alternative “Bag of Words” sentence level language understanding approach. It assigns sentences to classes that are associated with sets of logical forms in the same representation language as CARMEL produces. Thus, output from either source is appropriate input for the DLU module. However, the classification approach has the drawback that it embodies the underlying simplifying assumption that students always express required points in a single sentence, which is not always the case. 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., 2002). 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. For our combined approach, we use a decision tree trained with the ID3 decision tree learning algorithm (Mitchel, 1997) to combine Rainbow’s prediction with syntactic information in order to formulate a hypothesis about the classification of each sentence. We extract syntactic features for each sentence from the representation constructed by the parser. These features 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. ID3 uses these features to construct a decision tree for identifying the correct classification of novel sentences.

5 Results

We conducted a series of experiments to evaluate our statistical, symbolic, and combined approach. We used as our test set a corpus of 33 essays collected during web-based tutoring sessions that were not used as development data. The web based tutoring sessions during which we collected this corpus involved university students and a human tutor where students were answering the question “Suppose you are running in a straight line at constant speed. You throw a pumpkin straight up. Where will it land? Explain.” We divided these essays into a total of 130 sentence segments. For 77% of the data, three difference coders hand-classified each segment as having one or none of the 6 points required to solve the essay problem. We computed a pairwise Kappa coefficient to measure the agreement between coders, which was always greater than .75. We then selected one coder to complete the coding of the remainder of the data. We used that coder’s data as a gold standard to use for measuring the performance of our alternative approaches. We computed average per essay performance over 25 trials of randomly selecting essays covering 10% of the corpus, training the decision tree using ID3 on the rest, and then testing the selected essays.

Since WHY2’s domain specific knowledge sources are early in their development, we expected the symbolic only approach to perform poorly, and it did. It got an analysis precision of 17%, recall of 19%, and false alarm rate of 33%. Averaged over the essays in our test set, this translates in to a KCD precision of 64%, recall of 78%, and false alarm rate of 81%. The statistical only approach performed better overall with an analysis precision of 75%, recall of 73%, and false alarm rate of 15%. This translates into an average KCD precision of 88%, recall of 90%, and false alarm rate of 27%. The combined approach performed best of all with an analysis precision of 88%, recall of 75%, and false alarm rate of 8%. Notice that the combined approach performs as well as or better than both the statistical and the symbolic approach on analysis precision, recall, and false alarm rate as well as KCD selection precision, recall, and false alarm rate. The most striking aspects of the results are that it achieves a 95% KCD recall, a full 5% increase over the statistical approach, which cutting the statistical approach’s analysis false alarm rate in half. The results for this combined approach are displayed in Figure 5.

6 Conclusions and Current Directions

In this paper we have discussed the problem of selecting appropriate Knowledge Construction Dialogues (KCDs) for the purpose of encouraging students to include important points in their qualitative physics explanations that are missing. We have presented a model for projecting analysis precision, recall, and false alarm rate into KCD selection precision, recall, and false alarm rate. We used this model to inform the design of a heuristic for combining predictions from a symbolic and a statistical approach to essay analysis. We have demonstrated that our combined approach outperforms both the symbolic and the statistical

approach alone in terms of both KCD selection precision, recall, and false alarm rate.

7 Acknowledgments

The authors would like to thank the rest of the Natural Language Tutoring group for their collaboration.

This research is supported by the Office of Naval Research, Cognitive and Neural Sciences Division MURI Grant N00014-00-1-0600 and NSF Grant 9720359 to CIRCLE, a center for research on intelligent tutoring.

References

- [1] V. Alevan, O. Popescu, and K. Koedinger. 2001. Pedagogical content knowledge in a tutorial dialogue system to support self-explanation. In *Papers of the AIED-2001 Workshop on Tutorial Dialogue Systems*.
- [2] C. Burgess, K. Livesay, and K. Lund. 1998. Explorations in context space: Words, sentences, discourse. *Discourse Processes*, 25(2):211–257.
- [3] M. Chi, N. de Leeuw, M. Chiu, and C. LaVancher. 1981. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3).
- [4] M. G. Core, J. D. Moore, and C. Zinn. 2001. Initiative management for tutorial dialogue. In *Proceedings of the NAACL Workshop Adaption in Dialogue Systems*.
- [5] M. S. Glass. 1999. *Broadening Input Understanding in an Intelligent Tutoring System*. Ph.D. thesis, Illinois Institute of Technology.
- [6] Pamela W. Jordan, Maxim Makatchev, Michael Ringenberg, and Kurt VanLehn. 2002. Engineering the Tacitus-lite weighted abductive inference engine for use in the Why-Atlas qualitative physics tutoring system. submitted.
- [7] W. Kintsch. 2002. Predication. to appear in the Cognitive Science Journal.
- [8] T. K. Landauer, P. W. Foltz, and D. Laham. 1998. Introduction to latent semantic analysis. To Appear in *Discourse Processes*.
- [9] Andrew Kachites McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. <http://www.cs.cmu.edu/mccallum/bow>.
- [10] Mitchel, T. 1997. *Machine Learning*. McGraw Hill.
- [11] C. P. Rosé, P. Jordan, M. Ringenberg, S. Siler, K. VanLehn, and A. Weinstein. 2001. Interactive conceptual tutoring in atlas-andes. In *Proceedings of Artificial Intelligence in Education*.
- [12] C. P. Rosé. 2000. A framework for robust sentence level interpretation. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*.
- [13] K. VanLehn, P. Jordan, C. P. Rosé, and The Natural Language Tutoring Group. 2002. The architecture of why2-atlas: a coach for qualitative physics essay writing. In *Proceedings of the Intelligent Tutoring Conference*.
- [14] P. Wiemer-Hastings and I. Zipitria. 2001. Rules for syntax, vectors for semantics. In *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society*.
- [15] P. Wiemer-Hastings, A. Graesser, D. Harter, and the Tutoring Research Group. 1998. The foundations and architecture of autotutor. In B. Goettl, H. Halff, C. Redfield, and V. Shute, editors, *Intelligent Tutoring Systems: 4th International Conference (ITS '98)*, pages 334–343. Springer Verlag.