Finding Meaningful Search Features for Automated Analysis of Short Responses to Conceptual Questions

Christopher M. Nakamura[1], Sytil K. Murphy[1], Michael Christel[2], Scott M. Stevens[2] and Dean A. Zollman[1]

[1]Kansas State University Physics Department, 116 Cardwell Hall, Manhattan, KS 66506
[2]Carnegie Mellon University Entertainment Technology Center, 700 Technology Drive, Pittsburgh, PA 15219

Abstract. The Pathway Active Learning Environment synthetic tutoring system is capable of collecting large numbers of students' short responses to open-ended questions. The analysis of these responses may provide insight into the utility of the system, as well as information about student understanding of physics. The free-response nature of our data lends itself to qualitative analysis, however large data sets benefit from automated analysis. Natural language processing and data mining approaches, such as clustering, have been of interest across a variety of fields for automating the analysis of qualitative data. However, content-specific vocabulary, an abundance of search features, some of which are irrelevant, and inherent limitations on computers' abilities to match meaning are challenges that must be overcome. In this paper we discuss an analysis protocol for training computer models for automated data analysis. The preliminary analysis of two sample questions is presented, demonstrating a baseline of success.

Keywords: data mining, machine learning, online learning, synthetic tutoring
PACS: 01.40.Fk

INTRODUCTION

The Pathway Active Learning Environment (PALE) is a synthetic online tutoring system we have developed to study how to best use various interactive and multimedia technologies to help students learn physics. The system is primarily composed of two parts: (1) video-based lesson activities and (2) a Flash video-based tutoring interface that can answer students' typed natural language questions with pre-recorded responses. The system logs students' queries to the synthetic tutor, their typed responses to questions and other interactions. Investigating patterns in the logs that gauge the system's effectiveness in different configurations is one of our goals. Establishing how to best make the system function like a tutor is another. Data mining, machine learning and natural language processing are broad and related sets of techniques that may be useful in these efforts.

In this paper we discuss our initial efforts to explore the utility of machine learning for automatic analysis of students' short (about 1-4 sentence) responses to open-ended questions from online physics lessons. The method we are investigating seeks to identify naturally emerging commonalities in the ideas students express, connect those commonalities to useful search features that a computer can identify and use those features to build robust models that can categorize future responses. The protocol can be used to identify features of student understanding for future research or provide automated feedback in online or large-enrollment classes.

There are two main obstacles that make it unclear whether this type of scheme will work in this context. The first is the length of the student responses. This type of scheme has been explored in biology education in the context of essay grading [1]. The larger quantity of text associated with essays allows for greater statistical confidence about the feature composition of each text item. Interest has however, been shown in using this type of analysis scheme on shorter blocks of text [2]. This has, to the best of our knowledge, not been done in Physics Education Research (PER). The second obstacle that we face stems from the size of our data sets. While ultimately we seek analysis methods that are scalable and appropriate for large data sets, researchers and teachers cannot always obtain large data sets. The data sets we analyze here are not very large by data mining standards. Larger data sets are, of course, better. More data provides statistical confidence in similarities or differences in feature composition between different text items. At the same time it is useful to determine whether this scheme will work on smaller data sets. If it does, we would have confidence that it is appropriate for larger sets. We would also have increased incentive to obtain larger data sets because we would have another tool appropriate for their analysis.

In this paper we focus on a preliminary analysis of student responses to two sample questions from our lesson on Newton's first law. The goal is to explore the feasibility of using this type of analysis on a large scale by first trying it on a small subset of data.

For this analysis we use a software utility developed for annotating text data called the Summarization IDE (SIDE) [3]. The central component of SIDE is a text analysis program called TagHelper that extracts search features from the data set and contains algorithms to train models [2]. The utility is free and can be downloaded [4].

DATA COLLECTION

The responses analyzed here were collected in the fall of 2010 (F10) and summer of 2011 (S11). The F10 data was collected from several classes. The instructor of a large-enrollment concept-based physics class for elementary education majors assigned the completion of the lessons for homework. There were 105 students enrolled in that class. Twenty-three student volunteers from an algebra-based mechanics class were recruited and paid a small compensation to complete the lessons in our interview facility. Two high school teachers also used the system. Their 41 students were assigned the completion of the lessons independently in class. In S11 the instructor of a small algebra-based mechanics course assigned the completion of the first two lessons for homework. Twenty-nine students in that class completed these lessons in a university computer classroom.

We recognize that these groups of students are different in many ways, but they also have similarities. Since our current goal is to look for thematic commonalities amongst as many responses as possible, it is appropriate to combine data from these three courses and analyze them together. We note that students used the system in different configurations with and without the video tutor and completed the lessons under different conditions. Our goal is not to compare responses from students in different experimental groups, but to find themes in responses. Therefore it is appropriate to group the responses that were collected under different circumstances.

DATA REDUCTION & ANALYSIS

Prior to analysis a few preparatory steps are required. Data is manually filtered to remove student responses that are irrelevant, inappropriate or in other ways not suitable for analysis. The remaining responses are checked for spelling both via software and visual inspection. This step is important because the analysis tools are not robust against typographical errors. The spell-checking is minimally invasive and preserves the meaning of student responses.

With the student responses prepared the analysis procedure follows the process outlined in Fig. 1. The first step is to compare student responses to a given question to each other to ascertain how the responses are most naturally grouped. This step is subjective but key; failure to identify meaningful groups will render the analysis useless. However, the analysis protocol may also act as a check on the groups and help researchers to go back and refine the groups to produce a more meaningful analysis. Responses are put into the same group if the students use similar words to address similar ideas. Nominally 10 groups have emerged in analysis of responses, thus far.

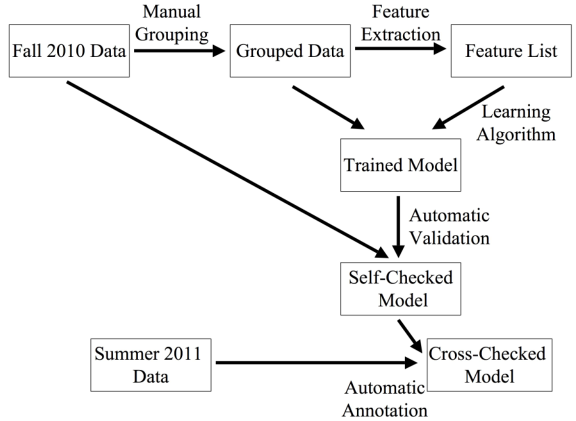

FIGURE 1. Student responses are grouped by similarities in the ideas expressed. Features are extracted and a model is trained. New data is used test the model.

Once the responses have been grouped they undergo computerized feature extraction. The features that are of interest in this analysis are words (unigrams), punctuation, response length and groups of words (n-grams). The SIDE utility can automatically extract all these features up to bigrams. If trigrams or larger features are desired another utility must be used. We have found that even bigrams increase analysis time significantly (they can triple the size of the feature list) without improving the results.

The feature list and the grouped data are then used to create a trained model via one of a number of classification algorithms. NaïveBayes and Support Vector Machines (SVM) are two that are appropriate in this case [2,3]. The SIDE utility automatically regroups the input data set using the trained model. This provides the first check on the quality of the model and the thematic coherence of the data. We then use the model to analyze an independent data set. Since this data was not used to generate the model, the successful analysis via the model provides greater insight into the utility of the model. Perhaps more importantly it provides greater confidence that the groupings that emerged from the first data might be of general use for analyzing future data.

Analysis of Question 1

In this question students were presented with the problem of a coin stuck inside a graduated cylinder and were asked to propose a method of obtaining the coin using Newton's first law. The method we had in mind is analogous to a car crash (which the students had seen in a previous activity) in that the coin and cylinder are brought into motion together and the cylinder is brought to a stop using a table, your hand or something similar and the coin continues its motion.

There were 122 analyzable responses to this question in F10. These were divided into nine groups as shown in Table 1. Several feature lists were extracted and tested. The best results were obtained when only unigrams were used. We have not exhaustively tested each feature's utility and we cannot say that our feature list is optimized. It represents the best currently obtained. Using this feature list and the original data we trained a model using SIDE's NaïveBayes classifier. Testing with SVM produced similar results, however, SVM was significantly slower on our machine. NaïveBayes was therefore used for this analysis.

The trained model correctly classified 86 of the 122 responses, corresponding to a success rate of 70.5%. The success and failure of the model on a group-by-group basis is contained in Table 1. It is worth noting that for large groups comprised of more than 20 responses the model does better than the overall 70.5% success rate which indicates matching at worst 74% of responses and 90% or better in the remaining groups. Applying the model to 28 responses collected in S11 resulted in 17 matches and a success rate of 61%.

TABLE 1. Group-by-Group Distribution of Responses for Question 1

Group Name	Number in F10 Dataset	Matched (Other Ideas Included)	Matched (Other Ideas Excluded)	Number in S11 Dataset	Matched
Invert Cylinder	27	25 (92.6%)	25 (92.6%)	4	4 (100%)
Force Must be Applied	22	21 (95.5%)	21 (95.5%)	2	1 (50%)
Hit Cylinder on the Table	20	18 (90.0%)	18 (90.0%)	6	5 (83%)
Hit the Bottom of the Cylinder	23	17 (73.9%)	18 (78.3%)	8	5 (63%)
Focus on Gravity	7	2 (28.6%)	2 (28.6%)	3	1 (33%)
Focus on Inertia	5	2 (40.0%)	3 (60%)	1	1 (100%)
Multiple Methods Proposed	5	0 (0%)	0 (0%)	4	0 (0%)
Shake it	3	0 (0%)	0 (0%)	0	-
Other Ideas (Misc.)	10	1 (10%)	-	0	-
Total	122	86 (70.5%)	87 (77.7%)	28	17 (61%)

An emerging challenge in this analysis is that while there is significant coherence within the response set, there is also a significant number of student responses that don't fit in with the others. There are several options for dealing with these responses. One can put each one in its own group, group them all together in a single group or one can delete them from the data set on the grounds that they may be noise in the data. Making each one its own group introduces higher opportunity for confusion without any obvious benefits. The other two options have complex advantages and disadvantages that should be discussed in a longer paper. We analyzed the data with those responses grouped together in a group called Other Ideas and with those responses deleted. When the responses were deleted the resulting model correctly classified 87 of the remaining 112 responses for a success rate of 77.7%. The group-by-group performance is shown in Table 1. We note that virtually all of the "gain" in this approach is from the reduction of the number of responses that are analyzed and suppressing the additional failure rate these responses introduce. The number of correctly classified responses increased by only one response. Applying this model to the S11 data yielded classifications that were essentially identical to those obtained with the Other Ideas responses left in place; the same 17 responses were correctly classified.

Analysis of Question 2

In this question students were asked to consider a coin sitting on top of a card that in turn sits on top of a beaker. In this canonical demonstration of Newton's first law the card is quickly removed allowing the coin to fall straight down. Prior to observing a video of this demonstration the students were asked to predict the trajectory of the coin when the card was removed and then explain the prediction using Newton's first law. Here we analyze their explanations of the predictions.

There were 114 analyzable F10 responses. These were divided into 11 groups. The groups are described in Table 2. Again the best results were obtained with a feature list of only unigrams and the model was trained using NaïveBayes. The trained model correctly classified 69 of the F10 responses and missed on 45, for a success rate of 60.5%. The group-by-group breakdown is shown in Table 2. When applied to the

S11 data the trained model matched the human annotation 39% of the time.

TABLE 2. Group-by-Group Distribution of Responses for Question 2

Groups	Number in F10	Matched	Number in S11	Matched
No Forces on Coin	19	14 (73.7%)	2	1 (50%)
Focus on Card's Speed	17	14 (82.4%)	5	4 (80%)
Focus on Gravity	16	13 (81.3%)	4	1 (25%)
Prior Experiences	13	11 (84.6%)	0	-
Focus on Coin's Inertia	13	6 (46.2%)	7	3 (43%)
Coin Moves with Card	12	6 (50.0%)	2	2 (100%)
Coin & Card are Separate	8	0 (0%)	4	0 (0%)
Other Causal Connections	7	5 (71.4%)	0	-
Only Describes Motion	5	0 (0 %)	0	-
Other Ideas (Misc.)	4	0 (0%)	4	0 (0%)
Total	114	69 (60.5%)	28	11 (39%)

Here there are only four responses in the Other Ideas category. It is unlikely that excluding those responses would significantly change the results.

DISCUSSION

There are several interesting, if unsurprising, findings in this analysis. The first is that the protocol is more successful with classifying responses that belong to larger groups. The analysis is likely more effective with data sets that break down into similarly sized groups of 20 or more responses. The second is that noise exclusion may be of low importance. Our analysis of question 1 indicates that adding potentially noisy responses to the data set does not degrade the classification of other responses, but just introduces additional responses that are not easily classified.

The small size of the S11 data set makes it difficult to draw strong conclusions about the general utility of the protocol. There are points of strong and poor agreement between the human and computer within the S11 analysis. The cause of poor matching is not yet clear, and must be investigated. Better model training or modifying the grouping scheme may improve the results. An item-by-item inspection of S11 matching does suggest that responses containing multiple ideas confuse the system and that overlap in vocabulary between the groups is a problem. In the S11 analysis of question 1 none of the four responses that contained multiple ideas were matched by the model, but each was coded in accordance with one of the ideas expressed. The results suggest further work with larger data sets and more questions is needed.

FUTURE WORK & CONCLUSION

The next natural step in this analysis is to feed these results back to determine if the manually generated groups can be improved. Applying the protocol to more questions is underway. Analyzing larger data sets should also be performed. The method could also be compared to fully automated clustering. Text clustering has seen occasional use in PER [5]. However, we have not seen it performed on responses to short-answer questions. That type of analysis presents a significant advantage in reduced annotation time, but allows for less human insight.

The first criterion for success in training the models will be confidence that the models agree with human raters a majority of the time. In order to provide feedback a higher standard is likely needed. Students may not be satisfied if the system correctly interprets their responses 60% or 70% of the time. Additional work is needed to ascertain the acceptable failure rate.

While our results have not yet demonstrated the described protocol to be effective for general analysis, they have demonstrated some initial success and strongly suggest that additional study is warranted.

ACKNOWLEDGMENTS

This material is based in part upon work supported by the National Science Foundation under Grant numbers REC-0632587 and REC-0632657. We acknowledge Bryan Maher and Josh Gross for their work on PALE.

REFERENCES

1. R. H. Nehm and H. Haertig, *J. Sci. Educ. Technol.* **20,** 1-18 (2011).
2. C. Rosé, Y.C. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger, and F. Fischer, *Computer-Supported Collaborative Learning* **3**, 237-271 (2008).
3. E. Mayfield and C.P. Rosé, [unpublished] http://www.cs.cmu.edu/~emayfield/SIDE-documentation.pdf (last accessed July 8, 2011)
4. http://www.cs.cmu.edu/~cprose/SIDE.html (last accessed July 8, 2011)
5. B. Sherin, *Behavior Res. Meth.*, 40(1), 8-20 (2008).