Still No Crystal Ball: Toward an Application for Broad Evaluation of Student Understanding

Piper Armstrong
Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

Part of the Physical Sciences and Mathematics Commons

https://scholarsarchive.byu.edu/etd/9637

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.
ABSTRACT

Still No Crystal Ball: Toward an Application for Broad Evaluation of Student Understanding

Piper Armstrong
Department of Computer Science, BYU
Master of Science

Evaluation of student understanding of learning material is critical to effective teaching. Current computer-aided evaluation tools exist, such as Computer Adaptive Testing (CAT); however, they require expert knowledge to implement and update. We propose a novel task, to create an evaluation tool that can predict student performance (knowledge) based on general performance on test questions without expert curation of the questions or expert understanding of the evaluation tool. We implement two methods for creating such a tool, find both methods lacking, and urge further investigation.

Keywords: NLP, educational application, evaluation, topic modeling
ACKNOWLEDGMENTS

Thanks to my family and my lab mates.
Table of Contents

1 In Preparation: Still No Crystal Ball: Toward an Application for Broad Evaluation of Student Understanding 1
Chapter 1

In Preparation: Still No Crystal Ball: Toward an Application for Broad Evaluation of Student Understanding

This manuscript has not yet been accepted for publication.
Still No Crystal Ball: Toward an Application for Broad Evaluation of Student Understanding

Piper Armstrong
Computer Science
Brigham Young University
piper.armstrong@byu.edu

Kevin Seppi
Computer Science
Brigham Young University
kseppi@byu.edu

Abstract

Evaluation of student understanding of learning material is critical to effective teaching. Current computer-aided evaluation tools exist, such as Computer Adaptive Testing (CAT); however, they require expert knowledge to implement and update. We propose a novel task, to create an evaluation tool that can predict student understanding of the contents of a corpus based on general performance on exam questions from that corpus without expert curation of the questions or expert understanding of the evaluation tool. We implement two methods for creating such a tool, find both methods lacking, and urge further investigation.

1 Introduction

Knowing what students do and do not understand is essential for good teaching, and therefore effective evaluation of students is critical.

Perhaps one of the most well-known methods of evaluation using computer tools is Computer Adaptive Testing (CAT) (Gershon, 2005). However, the need for expert creation and curation of test items makes CAT expensive and impractical for everyday use (Mills and Stocking, 1996). Indeed, much current research seeks to solve this problem by automating the expert’s role of test item curation (Kurdi et al., 2020).

These limitations, which are, admittedly, being currently researched, prevent CAT tests from becoming ubiquitous learning tools usable at an individual or classroom level, though they are extremely useful for large groups and at the level of standardized testing.

CAT tests are also limited in that they require a single dimension to be defined, specifically question difficulty (Gershon, 2005). There are methods for creating multidimensional CAT tests, allowing a single test to be administered to evaluate more than one domain of knowledge (e.g. history and civics) however, these require even more expert involvement to implement and are still limited to the dimension of difficulty (Piton-Gonçalves and Aluí-sio, 2012). This single-dimensionality gives a very accurate but narrow view of student’s knowledge ranking as compared to their peers, rather than a broad view of individual strengths and weaknesses.

Topic models, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Anchor Words (Arora et al., 2013), are algorithms designed to automatically discover topics in a collection of documents based on word cooccurrences in those documents. A topic model is multi-dimensional, finding many topical relationships between texts. If a topic model could be used to evaluate understanding, it would bring the strength of that multidimensionality to the evaluation.

A computer-aided system that could dispense with the need for expert curation, and that could evaluate students without limitation to a single, pre-defined dimension would be invaluable in teaching at an individual level. Thus an instructor could have an automated tool at the beginning or middle of a course to determine the strengths and weaknesses of their students in order to tailor future teaching to meet specific needs.

We therefore propose a novel task to create a tool capable of multi-dimensional evaluation of student understanding of an underlying corpus.

This tool should be able to ask students questions from a logically connected corpus (i.e. common sense suggests that performance on history questions would not predict performance on math questions, so they should not be included in the same corpus); the tool should then use student responses to predict their understanding of the rest of the corpus, and be able to communicate what the student does and does not know to an instructor.

All data, models, and code required to replicate this study are available¹.

¹URL Redacted
Figure 1: A diagram of the proposed task. An exam is created from questions selected from a corpus of data. A student answers the selected questions, and those questions along with the original corpus are used to train a Model. The Model is then used to predict the student’s understanding of the text in the corpus. Our current research ends with these predictions; however, truly successful completion of the task should include a further step in which the Model produces an analysis or explanation of the student’s understanding for the instructor.

2 Methods

A diagram of the proposed tool is shown in Figure 1. The proposed tool would consist of the following major pieces: a dataset consisting of a corpus with questions, a model for prediction using student-provided labels (i.e. student responses to exam questions from the corpus), and a mechanism for communicating the meaning of the predictions to the instructor.

In this work, we focus on the first two elements of this tool. We use a dataset of textbooks, select questions from it randomly, and collect answers and labels from a user study. We create two topic models using our textbook data, and train two classifiers with the dataset, user labels, and topics.

Since our models are not sufficiently successful in fulfilling the first two elements of the proposed tool, we do not implement an analysis and explanation. However, a truly successful tool would need to successfully communicate an analysis or explanation of the student’s understanding to the instructor and so we briefly consider possible means of implementing such an explanation in the Models section of this research.

2.1 Dataset

The first requirement for the desired tool is a corpus of documents including questions that can be attempted by students. We also require student-provided labels or responses to exam questions.
2.1.1 Corpus and Exam Questions

OpenStax is an educational initiative started by Rice University to provide free peer-reviewed college textbooks to students (Stafford and Flatley, 2018). This collection of textbooks is available online and can be accessed in both an online, interactive format and a downloadable Portable Document Format (PDF). The collection includes textbooks on science, math, social sciences, etc. This resource is well-suited for the purposes of the study, as the textbooks include exercises.

We use the 14 English-language science textbooks, including each paragraph, exercise (or exercise and solution if available), and glossary term as a separate document in a corpus. The corpus contains approximately 81,000 documents with an average of roughly 100 words per document.

2.1.2 Labels

To collect student answers to exam questions, we conducted a user study approved by our institution’s ethics review board, recruiting 30 participants using flyers and snowball sampling. The participants were 18 to 53 years-old with median age of 22. There were 15 male and 15 female participants, and participants came from a wide range of backgrounds in science from those who had last taken a science class a decade previously, to current PhD candidates in a scientific field. The participants were randomly assigned to 3 groups, with 10 participants in each group. Each group spent an hour on the study.

The first group was shown random documents from the corpus and asked whether they knew the content of the document. The chose one of three labels: “I Knew This,” “I Did Not Know This,” or “Not Enough Information.” We call this group (Explicit), since they explicitly told us whether they knew the information presented or not.

The second group was given random exercises from the corpus and asked to complete the exercise, or, if there were multiple exercises in the documents, to complete the first exercise given. They could answer in a text box or click a button labeled “Not Enough Information” if there was not enough information to complete the exercise. They were also allowed to leave the exercise blank if they did not know how to complete the exercise. We call this group (Implicit) since their understanding of the underlying document was implied by whether they answered the question right or wrong.

The third group of participants (Hybrid) were given a combination of the above tasks. We wished to see whether what participants said they knew (explicit labels) could be used to predict their actual performance on questions (implicit labels). However, because of the poor performance of our models, we ultimately split out this group’s responses, and included their explicit and implicit responses with the responses of the other groups.

For the purposes of this study, the explicit labels—where participants were asked directly about their knowledge—were used as-is.

In an ideal world, the proposed tool would be capable of grading student responses, automatically turning them into labels. Unfortunately, the answers provided when participants attempted exercises from the corpus required hand-grading by our researchers. Many of the answers to the exercises were available directly in the document itself (though hidden from the participant). For those that were not available in the document or elsewhere in the corpus, the researchers used the textbook and if possible, other online sources, to verify the correctness of the answers given and labeled the answers as “I Knew this” if the participant correctly answered the question or “I Did Not Know This” if they answered it incorrectly. There was no attempt to verify that exercises marked as “Not Enough Information” qualified for this designation, and in all cases we ignored these labels. Exercises left blank were marked as “I Did Not Know This.”

Table 1 shows the average number of labels provided by each group. Participants labeled 37 to 643 documents with a median of 145 labels.

<table>
<thead>
<tr>
<th>Group</th>
<th>True</th>
<th>False</th>
<th>Lack Info</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implicit</td>
<td>23</td>
<td>146.1</td>
<td>12.7</td>
<td>181.8</td>
</tr>
<tr>
<td>Explicit</td>
<td>87.6</td>
<td>146.7</td>
<td>15.4</td>
<td>249.7</td>
</tr>
<tr>
<td>Hybrid</td>
<td>37.8</td>
<td>73.9</td>
<td>16.6</td>
<td>128.3</td>
</tr>
</tbody>
</table>

Table 1: The average number of labels provided by participants in the study for each label type: “I Knew This” (True), “I Did Not Know This” (False), and “Not Enough Information” (Lack Info). And the average total.

2.2 Models

The second requirement for the desired tool is a model capable of predicting student understanding of the underlying corpus.

For each model, we employ a topic model. We conjecture that the topical output from a successful predictive model could help convey the meaning
of the model’s output to an instructor, providing a starting point for the analysis and explanation—the third vital element of a successful tool.

We train the models using the textbook data, labels collected in the user study described above, and the topics from a topic model. The models predict on a held-out portion of user-study labels. We implement two models described below.

We used labels generated by one of our researchers answering exam questions for an hour (similar to the Implicit study group) to do a parameter sweep for each topic model to decide how many topics to use. We use 20 topics with tokens allowed to appear in 150 documents to 10500 documents.

2.2.1 Labeled Anchors

Labeled Anchors is a classifier built as an extension of the topic model Anchor Words (Arora et al., 2013; Lund et al., 2018). Anchor words is touted as being faster than more traditional models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Labeled Anchors essentially treats the classification of documents as an extra topic.

If Labeled Anchors could be shown to be an effective classifier for this tool, there would be several benefits. Although we do not expect expert curation of questions, Instructors should know what topics their students are expected to learn, and this knowledge could potentially be used to improve the model’s ability to predict student understanding. Research has shown that adding a human to the loop can improve labeled anchors performance on classification tasks. (Lund et al., 2018)

We use 5-fold cross-validation, and train adding and predicting on every label until we reach the 80% mark.

2.2.2 T5 Transformer

The T5 Transformer is a language model specifically built to explore transfer learning (Raffel et al., 2019). Built on a language model, it has been used for tasks from translation to sentiment classification, etc. In context of the current research, transfer learning could be invaluable because each student can only provide a small number of labels, not enough to train an entire model, but perhaps enough to fine-tune a pretrained model.

We take advantage of the T5 Transformers natu-
Figure 3: Each individual graph shows the accuracy achieved on a specific participant’s Implicit labels, i.e. they attempted exercises. The label in the bottom left corner of each chart identifies the participant whose labels are being predicted in that chart. The flat line is the Max Class baseline (the accuracy if we assume they will get everything right or everything wrong, whichever is higher), the other line is the Labeled Anchors model accuracy, and the point is the accuracy of T5 Transformer fine-tuned on a particular participant’s labels. The more blue or red a model is, the better or worse, respectively, it performed compared to the baseline.

We also require the model to predict the most likely topic as provided by Latent Dirichlet Allocation (Blei et al., 2003). In this case, the topic model is incidental to the performance of the classifier, but in future the topic could be useful for analysis and explanation to an instructor. We then fine-tune on the labels of the target study participant. We use 5-fold cross-validation. And conduct this training for each individual study participant.

3 Results

We compare the accuracy of predicting student labels when using Labeled Anchors and the T5 Transformer with the accuracy of predicting student labels if we assume all the student’s labels will be either “I Knew This” or “I Did Not Know This,” but favoring whichever gives a higher accuracy; we call this Max Class. Table 2 shows the average accuracy across each type of participant response, i.e. Explicit or Implicit. In each case, the Max Class baseline outperforms our models.

Figure 2 and Figure 3 show the accuracy achieved for each individual participants’ Explicit and Implicit labels respectively. For Explicit labels in 11 out of 20 attempts, one or both of the models outperformed the baseline. For Implicit labels in 6 out of 20 attempts, one or both models outperformed the baseline. No single model outperformed the baseline in more than 40% of instances.

Table 2: The average accuracy when predicting student labels. Max Class is the baseline and performs better on average than either of our other models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Implicit</th>
<th>Explicit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Class</td>
<td>0.79</td>
<td>0.66</td>
</tr>
<tr>
<td>Labeled Anchors</td>
<td>0.72</td>
<td>0.63</td>
</tr>
<tr>
<td>T5 Transformer</td>
<td>0.75</td>
<td>0.60</td>
</tr>
</tbody>
</table>
A tabular representation of the results is available in Appendix A.

4 Discussion

Both of our models performed worse than the max class baseline suggesting that predicting what a student knows based on other labels is a hard problem. Especially surprising are the poor results of predictions even using explicit labels; what—if not the document’s topic—are participants using to decide to mark a document as something they knew or did not know?

Perhaps paradoxically, the poor performance on the explicit label task is encouraging since it suggests that there may be something beyond a simple topical distinction that the current models are failing to capture and which future models may capture. If the problem were simply one of modality, we would expect to see larger label sets lead to better results. This is not the case. Indeed, as per Figure 3, Labeled Anchors performed best on a label set with only 44 labels, and the T5 Transformer performed best on a label set with only 18 labels. The handful of successes lend some hope for finding such a successful model.

Perhaps some curation is necessary. This could mean finding a mechanism other than random for choosing which document to present to the student, capitalizing on previous student responses. It could mean using a less diverse group of study participants (e.g. a group of students from a single biology class) leading to more effective transfer learning. It could mean incorporating questions from an instructor’s actual curriculum into the corpus and labeling and predicting on those questions.

5 Conclusion

We attempted to create a tool to evaluate students’ broad understanding of an underlying corpus without using an expert. Such a tool would be very useful.

We’re not there yet.

References


Table 3: The accuracy of predictions for the Max Class Baseline (Max), Labeled Anchors (LA), and T5 Transformer (T5)

A

Table 3 shows the final accuracy for each model on individual participant label sets.