Text Classification of Student Self-Explanations in College Physics Questions

Sameer Bhatnagar
Polytechnique Montreal

Michel Desmarais
Polytechnique Montreal

Nathaniel Lasry
John Abbott College

Elizabeth S. Charles
Dawson College

ABSTRACT

This study looks at the text data generated from the Asynchronous Peer Instruction tool, DALITE. The goals of this work are two-fold: i) to determine whether the words students use in their self-explanations can be predictive of their success on the related multiple-choice item, or even reveal their uncertainty about the concept being tested; and, ii) to determine if the collection of words used by a student over the course of a semester using DALITE can predict their end-of-semester learning outcomes. Through the course of this study, we examine the effectiveness of different statistical models and document representations to explain these data. Results suggest the following: i) words alone are not enough to reliably predict item-level outcomes, and ii) DALITE holds the potential to offer teachers a novel platform for formative assessment that is predictive of student success and learning.

Keywords

Natural Language Processing

1. INTRODUCTION

The Distributed Active Learning Integrated Technology Environment (DALITE)[2], implements an original peer instruction paradigm that relies on students providing a rationale to their choice over multiple-choice questions (MCQ). After every MCQ, the student is prompted to provide the rationale for their choice. Once provided, the student is shown a few other students' rationales for the same choice, and for an alternate choice. If the answer was right, the alternate choice shown is for a wrong answer, else it is the right answer's rationales. The student can then decide to change their choice or not.

This instruction paradigm has recently been integrated into the EdX platform and we believe it has a great future in MOOCs and other environments where educational crowdsourcing bootstraps instructional content. However, for the bootstrap to be effective, a good understanding of the process of learning from this type of content is crucial. This paper reports on early analysis of student rationales with this aim in mind, using a text classification framework.

For this particular study, we are interested in

- identifying students who are unsure about their answers (as revealed by when they switch from right-to-wrong, or wrong-to-right in DALITE). Are there linguistic patterns for students who are uncertain?
- studying the effect of the teacher on the development of their students' language. Is there a teacher effect?
- documenting group differences in language use, for subpopulations such as strong vs. at risk students, or male vs. female. [10] discusses the gender gap in performance in college physics classrooms. This was observed in a previous study of ours looking at DALITE as well[1]. Is there a measurable difference between the language used by strong students and weak ones? Are there gender differences?
- finding minimally disruptive, low-stakes, language based predictors of student failure, as early in the semester as possible. Can the results of DALITE questions assigned prior to any of the three midterms predict which students ultimately fail?
- which classification algorithms perform the best in this context? What document representations optimize classifier performance for the different target variables?

2. RELATED WORK

Peer Instruction[4], massively popularized by Eric Mazur from Harvard, encourages teachers to promote student discussions after question polls in class, and subsequently repoll their students to see if they changed their original answer choices. This is the process our group makes asynchronous using DALITE.

The classification of student texts in physics has been studied in work surrounding the Why2Atlas[18] intelligent tutoring system (ITS), which dialogues with students in order to lead them to a more complete conceptual understanding. [15] used deep syntactic analysis for the classification of statements as related to different components of a complete explanation of a concept. AutoTutor[7] and iStart[13] are conversational ITS' that use Latent Semantic Analysis[5] to analyze the coherence of student self-explanations. A recent review of advances in ITS's[16] showed that natural language processing techniques, such as distributional semantics and deep syntactic analysis could be used to classify speech acts by students.

The study described herein take the first step in analysing the data from the DALITE platform, which is not an ITS, and does not have any dialogue agents, but has students asynchronously interacting with each other's self-explanations. This first step is based on classifying students based on their words alone. Also, this study is slightly different than those mentionned above, as it was part of a larger in-vivo design based experiment on the development of tools to engage students in active learning, even outside of the classroom[3].

3. DATA AND METHODS

3.1 Corpus Statistics

The dataset is made up of student-generated self-explanations for 80 different DALITE items (conceptual physics questions). On average, 97 students attempted each item, writing explanations for each question with an approximate length of 32 words, with a type-token ratio of 0.87. The average number of unique words used by all students to answer any given one item was 310.

The 140 students in this study came from three different colleges in the province of Quebec, Canada. The course material was surrounding what would normally be freshman physics in the U.S. Besides collecting midterm grades and final course grades, each student also completed the Force Concept Inventory[8], at the beginning of the term, as well at the end. The normalized pre-post gain (or Hake gain) on this questionnaire has become a standard measure in the physics education research community. More aggregate statistics of the dataset rest are more fully described in [1].

3.2 Experimental Setup

In order to evaluate the classification models and document representations, we choose different target variables for each experiment:

1. Whether the rationale was written for a correct answer, or an incorrect one. This is meant as a starting baseline, just to make sure the texts are differentiable along this seemingly obvious axis (assuming students will use different words to justify different answer choices)

2. Whether the rationale was written by a student who ended up switching their answer choice for that question

3. Whether the rationale was written by a male student or a female student.

4. Which of the five teachers was the one who taught the student who wrote the rationale under examination.

High accuracy prediction by classifiers for these target variables might imply that DALITE could become more adaptive, modeling the student's cognitive state and future success, based solely on the words they used in their explanation alone. For our first set of experiments, feature matrices are built from ”per-question corpora” (one classifier is built for every question item in DALITE, so there is one document per student who attempted that question).The overall effectiveness measure used is accuracy, which is averaged across all per-question corpora. Classifier performance is always compared to a baseline accuracy, calculated by always predicting most frequent class for the target variable (reported at the top of each table in the results).

In addition, the research team also has the objective of providing teachers with early predictors of student failure. We approach this by building classifiers for a different set of target variables:

1. Will the identified student fail their first of three midterms?

2. Is this student at risk of failing the course (final grade within 5 percentage points of failing)?

3. Will this student end the term with a lower than median gain in conceptual knowledge? (as measured by the Hake gain on the FCI)

Building models that can discriminate between such students, based only on the words a student used in DALITE, could indicate whether or not that this tool offers teachers a valid and reliable type of formative assessment. For this second set of experiments, feature matrices are built from ”per-student corpora”, where one classifier is built to predict each target variable, and each data point is a concatenation of all the rationales written by the student up to a certain point in time (as we wish to see how early in the term a student's future troubles can be flagged).

3.3 Tools

All texts were automatically corrected using the PyEnchant module (any misspelled word was replaced by the most likely suggestion). For the experiments where we explore Part-of-Speech tagging, tokenized sentences were passed straight through the NLTK Part-of-Speech tagging module. All feature matrices and classifiers were built through the Scikit Learn library[14]. Classifier hyper-parameters were left at their default values. We employ Laplace smoothing for unseen words. In an effort to report generalizable results, we use stratified k-fold cross-validation, which preserves the class distribution in each fold.[1]

3.4 Statistical Models

Significant amount of work was done in comparing different statistical learning algorithms for text classification. One of the simplest yet most effective text classification approaches is the Naive Bayes classifier[11]. In datasets when vocabulary size was small, [12] compared different event models for

[1] all scripts used to get the results for this study are available at sameerbhatnagar.github.io/

the Naive Bayes family of classifiers, finding that the multivariate Bernoulli model (where the components of each document vector are binary, modeling simply the presence or absence of a word), performed better for text classification than its multinomial counterpart (where document vectors are the counts of the different terms in tha document). [9] shows that Support Vector Machines (SVM) are well suited to the task of text classification, due to three factors inherent to the nature of the task: high dimensional feature space, many relevant features (dense concept vectors), but sparse document vectors. Finally, we explore the utility of a k-nearest neighbor classifier in this setting as well, based on the intuition that the document vectors might not be linearly separable.

3.5 Document Vector Representations

This study also aims to explore different choices of document representation. The most basic choice would have the elements of document vectors simply containing raw word counts (we ensure that the words in the original questions item text are always included in the term-document matrices).[17] showed that shifting importance to rarer words across a corpus would improve classifier effectiveness. We also look at N-grams to relax the independence assumption between words, but this may require more data than we have to avoid sparsity (we only go up to bigrams). There is an interest in also adding syntactic information, such as part-of-speech (POS) tags, and represent documents as bags of POS-tags (e.g. since there is an important difference in physics between using the word "force" as a verb or as a noun, which could reveal a misconception if students use it incorrectly). Finally, document vectors can also be represented for their semantic content. One of the most successful techniques for this is Latent Semantic Analysis[5], which relies on a truncated singular value decomposition of term co-occurrence matrices. This allows us to approximately represent documents in a lower dimensional space, and typically removes noise such that document vectors that are similar in meaning, cluster together. The sensitive choice in such latent factor models is the choice of how many factors will be kept after the matrix decomposition. We do a grid search over different possible number of dimensions to reduce to, ranging from 2 to 10, and pick the model that performs best in cross-validation.

The total number of unique words used, which will serve as features for our models, is on the order of two times the number of students (the number of data points), resulting in term-document matrices that are very sparse (often more than 97%). For this reason we implement a univariate feature selection using a chi-squared test, where we select the top 10 ranked features for their usefulness in discriminating the target labels[11]. We experiment with different document representations, searching for the best combination.

4. RESULTS

Referring to the experimental setup section above, there are two sets of experiments that were conducted:

1. Given an item, and all the DALITE rationales written by the students in the past for that item, can we look only at the words entered by some new student, and predict his/her outcome on that item? Can we predict if they are going to switch their answer (and hence they might have been uncertain)? Can we predict their gender, or teacher?

2. Given a student, and all of the rationales written by that student up until a certain point in time, can we predict end-of-semester learning outcomes? Can we predict such outcomes early on in the semester?

Table 1: Accuracy for Predicting low hake gain for a student at end of term (Baseline accuracy: 0.52)

Model	Raw	TfIdf	Bigrams	PoS	SVD.r	SVD.B
Mult.NB	0.67	0.54	0.67	0.65	NA	NA
Bern.NB	**0.69***	**0.70***	**0.69***	0.59	NA	NA
SVM	0.64	0.51	0.64	0.48	0.51	0.51
kNN	0.63	**0.68***	0.65	0.45	0.55	0.55

Table 2: Accuracy for Predicting at risk for a student at end of term (Baseline accuracy: 0.56)

Model	Raw	TfIdf	Bigrams	PoS	SVD.r	SVD.B
Mult.NB	0.60	0.55	0.62	0.61	NA	NA
Bern.NB	0.62	**0.71***	0.64	0.63	NA	NA
SVM	0.55	0.55	0.54	0.55	0.55	0.55
kNN	0.62	**0.68***	0.53	0.65	0.61	0.63

We do not explicitly report in this paper any results from the first set of experiments, except in saying that none of our statistical models, with none of the possible document vector representations, was able to achieve a prediction accuracy above baseline.

We report a subset of the results for the second set of experiments in the tables below. Tables 1 and 2 show classifier performance if we include all of each student's rationales in making a prediction, while table 3 reduces the data available to the model to only include rationales written by students in the first third of the course. The rows of each table represent which statistical model was used, while the columns represent the choice of document vector representation ('SVD.r' means that truncated singular value decomposition is carried out on the feature matrix derived from the corpus of raw text, while 'SVD.B' means that that SVD was carried out on the feature matrix derived from the combinations of unigrams and bigrams. Since SVD return vectors with negative components, the Naive Bayes models are not applicable.)

5. DISCUSSION

Table 3: Accuracy for Predicting low hake gain for a student after 1/3 of semester (Baseline accuracy: 0.52)

Model	Raw	TfIdf	Bigrams	PoS	SVD.r	SVD.B
Mult.NB	**0.63***	0.59	**0.64***	0.62	NA	NA
Bern.NB	0.58	0.55	**0.66***	**0.65**	NA	NA
SVM	0.62	0.51	0.60	0.57	0.51	0.51
kNN	0.59	0.58	0.57	0.51	0.54	0.55

Our research team started this study with the following question: do students in different cognitive states, use different words to explain their thinking when answering conceptual questions? In general, the poor performance of most of the statistical models studied herein tends to confirm the intuition behind the body of work centered around Latent Semantic Analysis: in most cases, the mere occurrences of the words is not enough to discriminate strong students from weak ones, and that such datasets can be too noisy and sparse. However, what is striking how the truncated SVD models also yielded essentially null results (even when optimized for the number of dimensions through cross-validation). This may highlight how the DALITE rationales are too short in length to have similar success to other educational text classification systems using LSA, such as that of [6], where the input texts were long enough to predict local coherence.

The inability of all these models to predict item-level outcomes, such as getting the answer correct, or whether a student is about to switch their answer, leads us to believe that richer syntactical and semantic representations will be required. Otherwise, *whether a student is male or female, certain of their answer or not, correct or incorrect in their answer choice, they cannot be distinguished by their words alone*. What is more, even if students are studying under the guidance of very different teachers, using different textbooks (as was the case in our study), they are still using similar language to defend their thinking in DALITE.

An encouraging result lies in the relative success of the Bernoulli Naive Bayes model in each of the three tables, which suggests that modeling the absence, as well as the presence of words is useful in predicting student learning outcomes. In table 3, both Naive Bayes models significantly increase prediction accuracy over baseline for distinguishing students who would have below median Hake gain on the FCI, *using only the words in their rationales in the first third of the course*.

6. FUTURE WORK

A more domain specific study is now underway, where we are looking at the types of DALITE items that had higher than expected prediction accuracies from their associated classifiers. What do these items have in common? What are the most informative features for these models? How can such information help teachers in instructional design?

The most important facet of DALITE that has not yet been studied lies in the patterns in student preferences: when students are on the page where they can read their peers' rationales, and are asked to reconsider their original answer choice, they are also prompted to *select which, if any, of their peers' rationales they thought was most convincing*. This 'crowdsourcing' of high quality, peer-assesed rationales if very healthy for the future of DALITE, but is also fertile ground for research related to the current study: what distinguishes language that is effective to convincing to students (whether for the right answer, or the wrong one)?

7. ACKNOWLEDGMENTS

We would like to thank the teachers who participated in this study, without whose support this work would not be possible: Chris Whittaker (Dawson College), Kevin Lenton (John Abbott College), and Kevin Lenton (Vanier College). This work was funded through *Programme de Recherche sur l'Apprentissage et l'Enseignement* from the government of Quebec. Finally, we remain indebted to all our students who actively contributed to DALITE through their participation.

8. REFERENCES

[1] S. Bhatnagar, M. Desmarais, C. Whittaker, N. Lasry, M. Dugdale, and E. S. Charles. An analysis of peer-submitted and peer-reviewed answer rationales, in an asynchronous peer instruction based learning environment.

[2] E. Charles-Woods, C. Whittaker, M. Dugdale, N. Lasry, K. Lenton, and S. Bhatnagar. Designing of dalite: Bringing peer instruction on-line. In N. Rummel, M. Kapur, M. Nathan, and S. Puntambekar, editors, *Computer Supported Collaborative Learning*.

[3] E. Charles-Woods, C. Whittaker, M. Dugdale, N. Lasry, K. Lenton, and S. Bhatnagar. Beyond and within classroom walls: Designing principled pedagogical tools for students and faculty uptake. In *Computer Supported Collaborative Learning (in press)*, 2015.

[4] C. H. Crouch and E. Mazur. Peer instruction: Ten years of experience and results. *American Journal of Physics*, 69(9):970–977, 2001.

[5] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[6] P. W. Foltz, W. Kintsch, and T. K. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307, 1998.

[7] A. C. Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, D. Harter, T. R. G. Tutoring Research Group, and N. Person. Using latent semantic analysis to evaluate the contributions of students in autotutor. *Interactive Learning Environments*, 8(2):129–147, 2000.

[8] D. Hestenes, M. Wells, and G. Swackhamer. Force concept inventory. *The physics teacher*, 30(3):141–158, 1992.

[9] T. Joachims. *Text categorization with support vector machines: Learning with many relevant features.* Springer, 1998.

[10] L. E. Kost, S. J. Pollock, and N. D. Finkelstein. Characterizing the gender gap in introductory physics. *Physical Review Special Topics-Physics Education Research*, 5(1):010101, 2009.

[11] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[12] A. McCallum, K. Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[13] D. S. McNamara, C. Boonthum, I. Levinstein, and K. Millis. Evaluating self-explanations in istart: Comparing word-based and lsa algorithms. *Handbook of latent semantic analysis*, pages 227–241, 2007.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel,

B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] C. P. Rosé, A. Roque, D. Bhembe, and K. Vanlehn. A hybrid text classification approach for analysis of student essays. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2*, pages 68–75. Association for Computational Linguistics, 2003.

[16] V. Rus, S. D'Mello, X. Hu, and A. Graesser. Recent advances in conversational intelligent tutoring systems. *AI magazine*, 34(3):42–54, 2013.

[17] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[18] K. VanLehn, P. W. Jordan, C. P. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, et al. The architecture of why2-atlas: A coach for qualitative physics essay writing. In *Intelligent tutoring systems*, pages 158–167. Springer, 2002.