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. Weak results suggest richer syntactic ans semantic models of text are needed.

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. [6] 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. DATA AND METHODS

2.1 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[4], 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].

2.2 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[7]. In datasets when vocabulary size was small, [8] compared different event models for 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 that document). [5] 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.

2.3 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).[9] 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[3], 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.

3. DISCUSSION

None of the results are presented here, due to space limitations.[1].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. 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.

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

4. FUTURE WORK

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)?

5. 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.

6. 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] 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.

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

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

[6] 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.

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

[8] 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.

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