Data-Mining Textual Responses to Uncover Misconception Patterns

Joshua Michalenko[1], **Andrew S. Lan**[2], **Richard G. Baraniuk**[1]
[1]Rice University, [2]Princeton University
{jjm7@rice.edu, andrew.lan@princeton.edu, richb@rice.edu}

ABSTRACT

An important, yet largely unstudied, problem in student data analysis is to detect misconceptions from students' responses to *open-response* questions. Misconception detection enables instructors to deliver more targeted feedback on the misconceptions exhibited by many students in their class, thus improving the quality of instruction. In this paper, we propose a new natural language processing (NLP) framework to detect the common misconceptions among students' textual responses to open-response, short-answer questions. We introduce a probabilistic model for students' textual responses involving misconceptions and experimentally validate it on a real-world student-response dataset. Preliminary experimental results show that our proposed framework excels at classifying whether a response exhibits one or more misconceptions. More importantly, it can also automatically detect the common misconceptions exhibited across responses from multiple students to multiple questions; this is especially important at large scale, since instructors will no longer need to manually specify all possible misconceptions that students might exhibit.

Author Keywords

Learning analytics, Misconception detection, Natural language processing

INTRODUCTION

The rapid developments of large-scale learning platforms (e.g., MOOCs (edx.org, coursera.org) and OpenStax Tutor (openstaxtutor.org)) have enabled not only access to high-quality learning resources to a large number of students, but also the collection of student data at very large scale. The scale of this data presents a great opportunity to revolutionize education by using machine learning algorithms to *automatically* deliver personalized analytics and feedback to students and instructors in order to dramatically improve the quality of teaching and learning.

L@S 2017, April 20-21, 2017, Cambridge, MA, USA.

http://dx.doi.org/10.1145/3051457.3053996

Detecting misconceptions from student-response data

The predominant form of student data, their *responses* to assessment questions, contain rich information on their knowledge states. Analyzing why a student answers a question incorrectly is of crucial importance to deliver timely and effective feedback. Among the possible causes for a student to answer a question incorrectly, exhibiting one or more *misconceptions* is critical, since upon detection of a misconception, it is very important to provide targeted feedback to a student to correct their misconception in a timely manner. Examples of using misconceptions to improve instruction include incorporating misconceptions to design better distractors for multiple-choice questions [8], implementing a dialogue-based tutor to detect misconceptions and provide corresponding feedback to help students self-practice [16], preparing prospective instructors by examining the causes of common misconceptions among students [15], and incorporating misconceptions into item response theory (IRT) for learning analytics [14].

The conventional way of leveraging misconceptions is to rely on a set of pre-defined misconceptions provided by domain experts [4, 8, 15, 16]. However, this approach is not scalable, since it requires a large amount of human effort and is domain-specific. With the large scale of student data at our disposal, a more scalable approach is to automatically detect misconceptions from data.

Recently, researchers have developed approaches for data-driven misconception detection; most of these approaches analyze students' response to *multiple-choice* questions. Examples of these approaches include detecting misconceptions in mathematics and modeling students' progress in correcting them [7] via the additive factor model [2], detecting misconceptions in chemistry by monitoring group discussions [12], and clustering students' responses across a number of multiple-choice physics questions [17]. However, multiple-choice questions have been shown to be inferior to open-response questions in terms of pedagogical value [6]. Indeed, students' responses to open-response questions can offer deeper insights into their knowledge.

To date, detecting misconceptions from students' responses to open-response questions has largely remained an unexplored problem. A few recent developments work exclusively with *structured* responses, e.g., sketches [13], short mathematical expressions [9], and algebra with simple syntax [3].

Contributions

In this paper, we propose a NLP framework that detects students' common misconceptions from their *textual* responses to open-response, short-answer questions. This problem is very difficult, since the responses are, in general, *unstructured*.

Our framework consists of the following steps. First, we transform students' textual responses to a number of short-answer questions into low-dimensional feature vectors using well-known word-vector embedding tools. The embedding is then input to a new statistical model that jointly models both the textual feature vectors and expert labels on whether a response exhibits one or more misconceptions; these labels identify only *whether or not* a response exhibits one or more misconceptions but not *which* misconception it exhibits. Our model uses a series of latent variables: the feature vectors corresponding to the correct response to each question, the feature vectors corresponding to each misconception, the tendency of each student to exhibit each misconception, and the confusion level of each question on each misconception. We develop a Markov Chain Monte Carlo (MCMC) algorithm for parameter inference under the proposed statistical model; details regarding this algorithm are omitted due to space constraints.

We experimentally validate our framework on two real-world educational datasets collected from two high-school classes, one on AP biology and one on high school Physics. Our experimental results show that our framework excels at classifying whether a response exhibits one or more misconceptions compared to standard classification algorithms. More importantly, we show an example of a common misconception detected from our datasets and discuss how this information can be used to deliver targeted feedback to help students correct their misconceptions.

STATISTICAL MODEL

We now detail the statistical model for textual responses and misconceptions; its graphical model is visualized in Figure 1. Concretely, let there be a total of N students, Q questions, and K misconceptions. Let $M_{i,j} \in \{0, 1\}$ denote the binary-valued misconception label on the response of student j to question i provided by an expert grader, with $j \in \{1, \ldots, N\}$ and $i \in \{1, \ldots, Q\}$, where 1 represents the presence of (one or more) misconceptions, and 0 represents no misconceptions.

We transform the raw text of student j's response to question i into a D-dimensional real-valued feature vector, denoted by $\mathbf{f}_{i,j} \in \mathbb{R}^D$, via a pre-processing step (detailed later in the experimental setup section). Let $\Omega \subseteq \{1, \ldots, Q\} \times \{1, \ldots, N\}$ denote the subset of student responses that are labeled, since every student only responds to a subset of the questions.

We denote the *tendency* of student j to exhibit misconception k, with $k \in \{1, \ldots, K\}$ as $c_{k,j} \in \mathbb{R}$, and the *confusion level* of question i on misconception k, as $d_{i,k} \in \mathbb{R}$. Then, let $P_{i,j,k} \in \{0, 1\}$ denote the binary-valued latent variable that represents whether student j exhibits misconception k in their response to question i, with 1 denoting that the misconception is present and 0 otherwise. We model $P_{i,j,k}$ as a Bernoulli random variable

$$p(P_{i,j,k} = 1) = \Phi(c_{k,j} + d_{i,k}), \quad (i, j) \in \Omega,$$

where $\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(t; 0, 1)\mathrm{d}t$ denotes the inverse probit link function (the cumulative distribution function of the standard normal random variable). Given $P_{i,j,k}$ $\forall k$, we model the observed misconception label $M_{i,j}$ as

$$M_{i,j} = \begin{cases} 0 & \text{if } P_{i,j,k} = 0 \ \forall k, \\ 1 & \text{otherwise,} \end{cases} \qquad (i, j) \in \Omega.$$

In words, a response is labeled as having a misconception if one or more misconceptions is present (given by the latent misconception exhibition variables $P_{i,j,k}$). Given $P_{i,j,k}$ $\forall k$, the textual response feature vector that corresponds to student j's response to question i, $\mathbf{f}_{i,j}$, is modeled as

$$\mathbf{f}_{i,j} \sim \mathcal{N}(\boldsymbol{\gamma}_i + \sum_k P_{i,j,k}\boldsymbol{\theta}_k, \boldsymbol{\Sigma}_F), \quad \forall (i, j) \in \Omega,$$

where $\boldsymbol{\gamma}_i$ denotes the feature vector that corresponds to the correct response to question i, $\boldsymbol{\theta}_k$ denotes the feature vector that corresponds to misconception k, and $\boldsymbol{\Sigma}_F$ denotes the covariance matrix of the multivariate normal distribution characterizing the feature vectors. In other words, the feature vector of each response is a *mixture* of the feature vectors corresponding to the correct response to the question and each misconception the student exhibits.

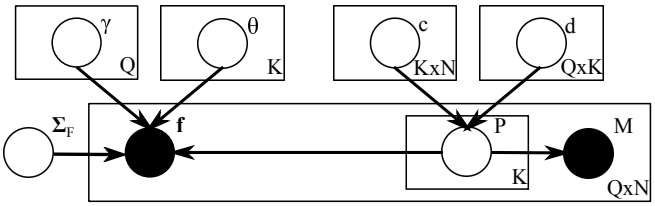

Figure 1: Visualization of the statistical model. Black nodes denote observed data; white nodes denote latent variables to be inferred.

EXPERIMENTS

We experimentally validate the efficacy of our proposed using two real-world educational datasets. We first detail the datasets, then compare the proposed framework against a baseline random forest (RF) classifier that classifies whether a student response exhibits one or more misconceptions. We conclude by showing a common misconception detected in our datasets and discuss how the proposed framework can use this information to deliver meaningful targeted feedback to students that helps them correct their misconceptions.

Dataset

Our two datasets consist of students' textual responses to short-answer questions in high-school classes administered on OpenStax Tutor [11] on two subjects: AP biology and Physics. For AP biology, $N = 113$ students responded to a total of 798 questions, and for Physics, $N = 208$ and $Q = 99$. Not every student responded to every question, which resulted in a total of 13131 responses in AP biology

and 1177 responses in Physics. Every response was labeled by an expert grader as to whether it exhibited one or more misconceptions.

Experimental setup

We first perform a pre-processing step by transforming each textual student response into a corresponding real-valued vector via word-vector embeddings. We train a standard skip-gram Word2Vec model [10] over the Openstax Biology and Physics textbooks (an approach also mentioned in [1]), to learn embeddings that put more emphasis on the technical vocabulary specific to each subject. We create the feature vector for each response by mapping each individual word in the response to its corresponding feature vector, and then adding them together. Concretely, denote the textual response of student j to question i, $\mathbf{x}_{i,j} = \{w_1, w_2, ..., w_{T_{i,j}}\}$ as the collection of words in the response, where $T_{i,j}$ denotes the total number of words in this response (excluding common stopwords). We then map each word w_t to its corresponding D-dimensional feature vector $r(w_t) \in \mathbb{R}^D$ using the trained Word2Vec model. We use $D = 10$ in our experiments. We then compute the student response feature vector as $\mathbf{f}_{i,j} = \sum_{t=1}^{T_{i,j}} r(w_t)$.

The assessment questions in AP Biology and Physics draw questions from the OpenStax textbooks; we divide the full AP Biology dataset into smaller subsets corresponding to each of the first four units of the Biology textbook , since different units correspond to entirely different sub-areas in biology. We do not further divide the Physics dataset since it is too small. We also trim each dataset by filtering out students who respond to less than 10 questions and questions with less than 10 responses in every dataset. We perform our experiments with $K = 2$ latent misconceptions.[1] We compare the proposed framework against a baseline random forest (RF) classifier[2] using the textual response feature vectors $\mathbf{f}_{i,j}$ to classify the binary-valued misconception label $M_{i,j}$, with 100 decision trees.

We randomly partition each dataset into 5 folds and use 4 folds as the training set and the other fold as the test set. We then train both algorithms on the training set and evaluate their performance on the test set, using two metrics: i) prediction accuracy (ACC), i.e., the portion of correct predictions, and ii) area under curve (AUC), i.e., the area under the receiver operating characteristic (ROC) curve of the resulting binary classifier [5]. Both metrics take values in $[0, 1]$, with larger values corresponding to better prediction performance. We repeat our experiments for 20 random partitions of the folds.

Results and discussion

We compare the performance of the proposed framework against RF on misconception label classification in Table 1 and Table 2. Our proposed framework significantly outperforms RF (1–4% using the ACC metric and 4-17% using the AUC metric). The performance gain of the proposed framework over RF is larger for the AP Biology datasets, in which students write longer textual responses and smaller in the Physics dataset due to the fact that most responses therein only contain a few words.

We emphasize that, in addition to the proposed framework's significant improvement over RF in terms of misconception label classification, it features great interpretability since it identifies common misconceptions from data. For example, the following responses from multiple students across two questions are identified to exhibit the same misconception in the AP Biology Unit 4 dataset:

> *Question 1*: People who breed domesticated animals try to avoid inbreeding even though most domesticated animals are indiscriminate. Evaluate why this is a good practice.
> *Correct Response*: A breeder would not allow close relatives to mate, because inbreeding can bring together deleterious recessive mutations that can cause abnormalities and susceptibility to disease.
> **Student Response 1**: Inbreeding can cause a rise in unfavorable or detrimental traits such as genes that cause individuals to be prone to disease or have unfavorable mutations.
> **Student Response 2**: Interbreeding can lead to harmful mutations.

> *Question 2*: When closely related individuals mate with each other, or inbreed, the offspring are often not as fit as the offspring of two unrelated individuals. Why?
> *Correct Response*: Inbreeding can bring together rare, deleterious mutations that lead to harmful phenotypes.
> **Student Response 3**: Leads to more homozygous recessive genes thus leading to mutation or disease.
> **Student Response 4**: When related individuals mate it can lead to harmful mutations.

Although these responses are from different students to different questions, they exhibit one common misconception, that inbreeding leads to harmful mutations. Once this misconception is identified, course instructors can deliver the targeted feedback that inbreeding only brings together harmful mutations, leading to issues like abnormalities, rather than directly leading to harmful mutations.

Moreover, the proposed framework can automatically discover common misconceptions that students exhibit without input from domain experts, especially when the number of students and questions are very large. Specifically, in the example above, we are able to detect such a common misconception that 4 responses exhibit by analyzing the 2278 responses in the AP Biology Unit 4 dataset; however, it would not likely be detected if the number of responses was smaller and fewer students exhibited the misconception. This feature makes it an attractive data-driven aid to domain experts in designing content to address student misconceptions.

[1] We omit experimental results with other values of K due to spatial constraints.

[2] The RF classifier achieves the best performance among a number of off-the-shelf baseline classifiers, e.g., logistic regression, support vector machines, etc. Therefore, we do not compare against other baseline classifiers.

Table 1: Performance comparison on misconception label classification of a textual response in terms of the prediction accuracy (ACC) of the proposed framework against a random forest (RF) classifier.

	AP Biology unit 1	AP Biology unit 2	AP Biology unit 3	AP Biology unit 4	Physics
Proposed framework	0.789 ± 0.014	0.774 ± 0.015	0.779 ± 0.019	0.887 ± 0.011	0.756 ± 0.034
RF	0.762 ± 0.019	0.735 ± 0.011	0.758 ± 0.017	0.873 ± 0.009	0.745 ± 0.031

Table 2: Performance comparison on misconception label classification of a textual response in terms of the area under the receiver operating characteristic curve (AUC) of the proposed framework against RF.

	AP Biology unit 1	AP Biology unit 2	AP Biology unit 3	AP Biology unit 4	Physics
Proposed framework	0.762 ± 0.027	0.758 ± 0.023	0.752 ± 0.020	0.774 ± 0.029	0.782 ± 0.045
RF	0.645 ± 0.025	0.676 ± 0.014	0.630 ± 0.024	0.604 ± 0.034	0.746 ± 0.042

CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a NLP framework for detecting and classifying common misconceptions in students' textual responses. Our experiments on two real-world educational datasets consisting of students' textual responses to short-answer questions show that the proposed framework excels at classifying whether a response exhibits one or more misconceptions. Moreover, we are also able to group responses with the same misconceptions into clusters, enabling the data-driven discovery of common misconceptions without input from domain experts. Possible avenues of future work include i) test other word-vector embeddings that take word ordering into account, i.e., embeddings that map responses "If X then Y" and "If Y then X" to different feature vectors, and ii) automatically generate the appropriate feedback to correct each misconception.

REFERENCES

1. Bhatnagar, S., Desmarais, M., Lasry, N., and Charles, E. S. Text classification of student self-explanations in college physics questions. In *Proc. 9th Intl. Conf. Educ. Data Min.* (July 2016), 571–572.
2. Cen, H., Koedinger, K. R., and Junker, B. Learning factors analysis – A general method for cognitive model evaluation and improvement. In *Proc. 8th. Intl. Conf. Intell. Tutoring Syst.* (June 2006), 164–175.
3. Elmadani, M., Mathews, M., Mitrovic, A., Biswas, G., Wong, L. H., and Hirashima, T. Data-driven misconception discovery in constraint-based intelligent tutoring systems. In *Proc. 20th Int. Conf. Comput. in Educ.* (Nov. 2012), 1–8.
4. Griffiths, A. K., and Preston, K. R. Grade-12 students' misconceptions relating to fundamental characteristics of atoms and molecules. *J. Res. in Sci. Teaching 29*, 6 (Aug. 1992), 611–628.
5. Jin, H., and Ling, C. X. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowl. Data Eng. 17*, 3 (Mar. 2005), 299–310.
6. Kang, S., McDermott, K., and Roediger III, H. Test format and corrective feedback modify the effect of testing on long-term retention. *Eur. J. Cogn. Psychol. 19*, 4-5 (July 2007), 528–558.
7. Liu, R., Patel, R., and Koedinger, K. R. Modeling common misconceptions in learning process data. In *Proc. 6th Intl. Conf. on Learn. Analyt. & Knowl.* (Apr. 2016), 369–377.
8. Maass, J. K., and Pavlik Jr, P. I. Modeling the influence of format and depth during effortful retrieval practice. In *Proc. 9th Intl. Conf. Educ. Data Min.* (July 2016), 143–149.
9. McTavish, T., and Larusson, J. Discovering and describing types of mathematical errors. In *Proc. 7th Intl. Conf. Educ. Data Min.* (July 2014), 353–354.
10. Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (Sep. 2013).
11. OpenStax Tutor. https://openstaxtutor.org/, 2016.
12. Schmidt, H. J. Students' misconceptions—Looking for a pattern. *Sci. Educ. 81*, 2 (Apr. 1997), 123–135.
13. Smith, A., Wiebe, E. N., Mott, B. W., and Lester, J. C. SketchMiner: Mining learner-generated science drawings with topological abstraction. In *Proc. 7th Intl. Conf. Educ. Data Min.* (July 2014), 288–291.
14. Tatsuoka, K. K. Rule space: An approach for dealing with misconceptions based on item response theory. *J. Educ. Meas. 20*, 4 (Dec. 1983), 345–354.
15. Tirosh, D. Enhancing prospective teachers' knowledge of children's conceptions: The case of division of fractions. *J. Res. Math. Educ. 31*, 1 (Jan. 2000), 5–25.
16. VanLehn, K., Jordan, P. W., Rosé, C. P., Bhembe, D., Böttner, M., Gaydos, A., Makatchev, M., Pappuswamy, U., Ringenberg, M., Roque, A., Siler, S., and Srivastava, R. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proc. 6th Intl. Conf. on Intelligent Tutoring Systems* (June 2002), 158–167.
17. Zheng, G., Kim, S., Tan, Y., and Galyardt, A. Soft clustering of physics misconceptions using a mixed membership model. In *Proc. 9th Intl. Conf. Educ. Data Min.* (July 2016), 658–659.