Enhancing the Automatic Identification of Common Math Misconceptions Using Natural Language Processing

Guher Gorgun[1(✉)] and Anthony F. Botelho[2]

[1] University of Alberta, Edmonton, AB T6G 2G5, Canada
gorgun@ualberta.ca
[2] University of Florida, Gainesville, FL 32611, USA
abotelho@coe.ufl.edu

Abstract. In order to facilitate student learning, it is important to identify and remediate misconceptions and incomplete knowledge pertaining to the assigned material. In the domain of mathematics, prior research with computer-based learning systems has utilized the commonality of incorrect answers to problems as a way of identifying potential misconceptions among students. Much of this research, however, has been limited to the use of close-ended questions, such as multiple-choice and fill-in-the-blank problems. In this study, we explore the potential usage of natural language processing and clustering methods to examine potential misconceptions across student answers to both close- and open-ended problems. We find that our proposed methods show promise for distinguishing misconception from non-conception, but may need further development to improve the interpretability of specific misunderstandings exhibited through student explanations.

Keywords: misconceptions · sentence-BERT · intelligent tutoring system · natural language processing

1 Introduction

Educators across learning contexts and domains rely on a range of content to assess students' knowledge and understanding of covered concepts. In the domain of mathematics, for example, as is the focus of this paper, it is not uncommon for teachers to assign homework and classwork in the form of problem sets composed of multiple interleaved types of problems [5]. Traditionally, these different types of problems include formats of multiple choice, fill-in-the-blank, and short answer questions, but may also include other types of questions such as drawing charts and graphs, as well as essay questions (though the latter is likely less common in the domain of mathematics). Prior works have described these different types of problems by distinguishing "close-ended" questions from "open-ended" questions (e.g., [1]); while the scoring of student answers to close-ended

N. Wang et al. (Eds.): AIED 2023, CCIS 1831, pp. 302–307, 2023.
https://doi.org/10.1007/978-3-031-36336-8_47

problems is relatively easy to automate with a matching procedure as there is usually a small number of correct answers and the variation in possible answers to open-ended responses makes this task much more difficult.

In the past, notable research in addressing student misconceptions has been limited to observing student work on close-ended questions through "bugs" and "common wrong answers" (CWA; [5]). While common wrong answers, or particular incorrect responses that are answered by a large proportion of students, can be helpful in understanding student misconceptions, student responses to open-ended problems may provide even greater insights. Teachers often rely on student open-ended work to understand the thought processes and strategies taken by students to find a solution. Therefore, open-ended responses could provide opportunities to identify misconceptions with greater precision.

Recent advancements in natural language processing (NLP) have resulted in the application of deep learning embedding methods such as Sentence-BERT [2], which has been used in educational contexts to identify sets of similar student answers to open response problems (e.g. the method described in [1]). Such methods may be used to identify common incorrect explanations.

This paper represents a proof of concept in using NLP to identify misconceptions through common wrong answers in open-ended explanations. To test the feasibility of our approach, we examine student answers to a single 2-part problem from the ASSISTments learning platform. This 2-part problem consists of a close-ended "Part A" followed by an open-ended "Part B" that prompts students to explain their solution to the preceding part. Through a set of exploratory analyses, we seek to address the following research questions:

1. What are the common answers that emerge when clustering student-written explanations for a single open-ended mathematics problem?
2. Do similar sets of common incorrect answers emerge when comparing across close- and open-ended components of a single mathematics problem?

2 Methods

To conduct our analyses, we select a single 2-part problem from ASSISTments from a large set of student log data collected between 2018 through 2022. From this large set, we identify a candidate set of problems where the problems have at least 2 parts (consisting of a close-ended, followed by an open-ended question) and the second part prompts students to explain their work to the preceding part. Within the system, by default, teachers must manually score open-ended answers on an integer scale from 0–4. We filter problems where there were fewer than 2 teachers who provided scores to students and include only problems where all 5 score values were present in the data. We further filter out any problems where the percentage of unique answers is larger than 75% and 5% for open-ended and close-ended responses, respectively (to identify problems where there is notable variance in student responses to evaluate our methods). From the set of candidate problems, we select the problem that contains the largest sample size. The text of the close-ended portion of this problem is depicted in Fig. 1.

Fig. 1. The close-ended problem prompts for the selected problem set. The problem was followed by an open-ended question prompting students to explain their work.

2.1 Identifying Common Wrong Answers

In the first step of our analysis, we use the student answers to the close-ended portion of the problem to extract the most common incorrect answers. As a "select all that apply" question type, as seen in Fig. 1, students are asked to select all of the values equivalent to the ratio of 12:3. As this type of problem is graded by the computer by matching each student's answer against the known correct answer, we identify all incorrect student answers. From this, we identify the set of unique student answers and calculate a simple frequency to measure each incorrect answer's commonality (see Table 1).

To identify common explanations for open-ended component of the problem set selected, we conduct a second analysis on the subsequent part of the problem. With the set of student textual explanations, we first cleaned the set of answers by removing any HTML tags and accented characters (that would not be recognized by most NLP models), and removed any empty student responses. With these, we utilized a pre-trained Sentence-BERT model (SBERT; [2]) which converts each answer into a 768-valued feature vector. The intuition of this and similar embedding methods is that it creates an embedding space where the distance of each textual sample to all others is correlated with the semantic similarity of the language (i.e. similar student responses should cluster closely together within the space). After generating these embeddings, we grouped sets of student answers by the teacher-given score to identify groups of similar answers within each score band.

We used the k-means clustering method [4] to identify clusters within each score band. Yet, due to the high dimensionality of the data, the k-means cluster-

Table 1. The correctness statistics and common wrong answers for the close-ended component of the problem set selected.

N Correct/ N Incorrect	Correct Answer	3 Most Common Wrong Answers	Count of Wrong Answers
1064 (37.7%)/ 1761 (62.3%)	4:1, 24:6, 1200:300	24:6, 1200:300	172 (10.2%)
		2. 4:1, 24:6	140 (8.3%)
		6:1, 4:1, 24:6, 1200:300	89 (5.3%)

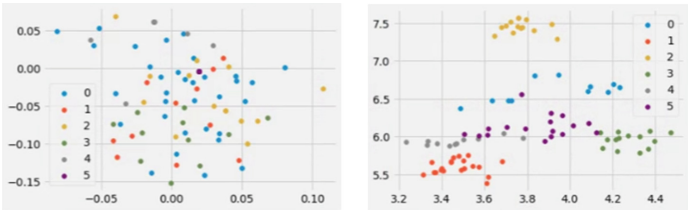

Fig. 2. Clusters identified before and after applying UMAP method.

ing method did not perform well as measured by the resulting clusters' silhouette score [7]. This score suggested poor coherence, indicating that it would be difficult to interpret meaningful differences between cluster groupings. Following a similar procedure to the BERTopic modeling algorithm [3], we applied a dimensionality reduction algorithm, Uniform Manifold Approximation and Projection (UMAP) [6] in an attempt to improve the clustering by removing redundant and irrelevant features from the embedding models (and simplifying the clustering procedure). The result of applying UMAP prior to k-means clustering rendered better clusters with higher values of silhouette coefficients and the sum of squared distances (SSD). This improvement can be seen in Fig. 2 which depicts the resulting cluster cohesion, through a 2D projection of the embedding space based on the most representative axes, before applying UMAP (left) and after (right). Finally, we identified the most frequent bi-grams present in each cluster after removing English stopwords. We also tried using unigrams and trigrams to identify keywords, however, they were not as helpful as using bi-grams in identifying the common theme of each cluster.

3 Results

We analyzed problems and clusters found for each score band in detail to identify common themes and misconceptions, as summarized in Table 2. Within each score band, we typically observed a cluster composed of students indicating that they did not know the answer. The only consistent exception was the score band with the full score where none of the students stated that they did not know the answer or made an accidental slip.

From examining the common wrong answers from the close-ended portion in Table 1, we see a large number of students failing to include the ratio of 4:1 and, to a lesser degree, missing the ratio of 1200:300 or including the incorrect ratio of 6:1. The first missing ratio could indicate that students struggled to represent the ratio in its simplest form; this could point to difficulties representing the ratio as a fraction or other errors when reducing that fraction. Similarly, the second-most-common wrong answer may suggest difficulties for students to identify the larger numbers as multiples of the given ratio. Finally, in the third CWA, the inclusion of the ratio 6:1 suggests a misunderstanding of what a ratio is meant to represent in terms of the relationship between the two numbers.

Table 2. The most frequent bi-grams observed in each cluster in Problem 1.

Score Band	N Samples	Keywords/Bi-grams in Clusters
0	53	Cluster 1: got wrong, didn't know, need help
	20	Cluster 2: idk idk, don't know, know idk
	14	Cluster 3: tp ratio, long ratios, use numbers
1	6	Cluster 1: kinda forgot, sorry got, problem sorry
	6	Cluster 2: scale factors, bc scale, copies multiply
	5	Cluster 3: 12 got, got right, little high
	4	Cluster 4: 1s numbers, added know, timesing numbers
	4	Cluster 5: picked divide, 12 picked, 24 12
2	12	Cluster 1: 12 didn't, 12 know, turn 12
	12	Cluster 2: different multiples, divided just, times different
	12	Cluster 3: don't wrong, got wrong, added don't
	8	Cluster 4: wrong just, knew 24, got wrong
	8	Cluster 5: fractions multiplied, fraction multiplied
	7	Cluster 6: fit 12, lower because, fit numbers
3	29	Cluster 1: number divide, multiply number, divided multiply
	15	Cluster 2: divide multiply, 12 24, factors ratio
	14	Cluster 3: times equals, 12 times, 12 multiplied
4	79	Cluster 1: multiply divide, know multiply, divide number
	64	Cluster 2: 200 300, 1200 300, 12 divided
	46	Cluster 3: 12 know, numbers 12, multiples 12
	36	Cluster 4: equal ratio, original ratio, ratio multiple
	30	Cluster 5: ratio 12, equal 12, equivalent 12
	28	Cluster 6: 12 times, times 100, 24 3×2

In comparing these to the clusters and bi-grams in Table 2, it is easy to first realize the large number of clusters containing "non-answers" such as "idk" and "don't know." It is interesting to also recognize that such clusters emerged in several score bands, suggesting that teachers provided some credit to students admitting their lack of understanding. While these types of clusters may indicate very little in terms of misconception, they are quite informative in identifying non-conception. In other words, it is difficult to ascertain from the close-ended problem which answers were deliberate and which were the result of a somewhat random selection. While it is easy to conclude that the inclusion of 6:1 might indicate a misunderstanding of the second number in a ratio or even the relationship between a numerator and denominator, it is also the case that this is

the first option provided to students and therefore it may just be the most likely first guess (e.g. the likelihood of students guessing answers is likely not uniform across the selections). Combining the close- and open-ended answers can help distinguish misconception from non-conception. In observing other clusters, we can deduce that some students are exhibiting at least partial knowledge by referencing keywords such as fractions, multiples, and division, but the bi-grams alone are seemingly not the most informative way of identifying specific misconceptions.

4 Conclusion

The analyses presented in this work contribute mixed results in terms of automating the identification of misconceptions in student textual explanations. We found that, although we are able to identify clusters of student answers, the use of bi-grams offers only limited utility in drawing conclusive interpretations as to what each cluster exhibits in terms of student understanding. With that, however, we found that the method we propose here is quite helpful in distinguishing between misconception and non-conception when taking into account the CWAs that emerge from close-ended problems.

Beyond this context, this work also offers contributions in the form of best practices when approaching the clustering of SBERT embeddings. We found that the use of UMAP was instrumental in producing clusters that were interpretable in any capacity. Future works attempting to identify sets of similar student answers should consider such methods to both simplify clustering procedures and improve the cohesion of resulting groups. Of course, as this work represents a simple proof-of-concept, future work is planned to scale these analyses to understand whether our findings generalize to larger sets of problems.

References

1. Baral, S., Botelho, A.F., Erickson, J.A., Benachamardi, P., Heffernan, N.T.: Improving automated scoring of student open responses in mathematics. International Educational Data Mining Society (2021)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Grootendorst, M.: BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794 (2022)
4. Hartigan, J.A., Wong, M.A.: Algorithm as 136: a k-means clustering algorithm. J. R. Stat. Soc. Ser. C (Appl. Stat.) **28**(1), 100–108 (1979)
5. Heffernan, N.T., Heffernan, C.L.: The ASSISTments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. Int. J. Artif. Intell. Educ. **24**(4), 470–497 (2014)
6. McInnes, L., Healy, J., Melville, J.: UMAP: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
7. Zhou, H.B., Gao, J.T.: Automatic method for determining cluster number based on silhouette coefficient. In: Advanced Materials Research, vol. 951, pp. 227–230. Trans Tech Publications (2014)