# Improving Automated Assessment and Feedback for Student Open-responses in Mathematics

Sami Baral
Worcester Polytechnic Institute
sbaral@wpi.edu

## ABSTRACT

Advancements in online learning platforms have revolutionized education in multiple different ways, transforming the learning experiences and instructional practices. The development of natural language processing and machine learning methods have helped understand and process student languages, comprehend their learning state, and build automated supports for teachers. With this, there has been a growing body of research in developing automated methods to assess students' work both in mathematical and non-mathematical domains. These automated methods address questions of two categories; closed-ended (with limited correct answers) and open-ended (are often subjective and have multiple correct answers), where open-ended questions are mostly used by teachers to learn about their student's understanding of a particular concept. Manually assessing and providing feedback to these open-ended questions is often arduous and time-consuming for teachers. For this reason, there have been several works to understand student responses to these open-ended questions to automate the assessment and provide constructive feedback to students. In this research, we seek to improve such a prior method for assessment and feedback suggestions for student open-ended works in mathematics. For this, we present an error analysis of the prior method "SBERT-Canberra" for auto-scoring, explore various factors that contribute to the error of the method, and propose solutions to improve upon the method by addressing these error factors. We further intend to expand this approach by improving feedback suggestions for teachers to give to their students' open-ended work.

## Keywords

Online Learning Platforms, Open-responses, Natural Language Processing, Machine Learning, Automated assessment, Mathematics

## 1. INTRODUCTION

In the past decade, development in artificial intelligence and machine learning methods have led to advancements in online learning platforms, transforming learning experiences and teaching practices. From personalized learning to augmenting teaching processes through automated assessment methods [8, 2, 4, 14, 11, 1], the benefits of these platforms have been significant. With this, several prior works have leveraged machine learning methods and natural language processing-based techniques to automate the assessment of students' work, both in mathematical [10, 6] and non-mathematical domains [13, 15]. As these methods and models of student learning become deeply integrated into normal instructional and educational practices, it becomes increasingly important to understand the strengths and weaknesses in their application. Within this, it is important to not only identify areas where existing methods under-perform, but it is also important to develop methods to improve such models to alleviate risks to fairness.

In the domain of mathematics there have been several works to automate assessment and provide constructive feedback to students, to further increase the efficiency of teaching and help teachers guide their focus to students in need. These works address problems of two categories: close-ended and open-ended problems. For close-ended problems, that have a finite number of correct answers, these auto-scoring methods can apply simple matching techniques to compare the student answer with the list of correct answers and consistently achieve near-perfect accuracy. However, open-ended problems are subjective with multiple accepted correct answers and are mostly given in the form of natural language. For these types of responses teachers commonly assess students based on an explicit or implicit rubric that pinpoints key aspects that must be included in a student response to sufficiently demonstrate their understanding. In addition, these types of student responses in mathematics often are a combination of language, images, tables, or other mathematical expressions, equations, and terminologies, which poses a challenge in developing automated methods of assessment for these problems. Due to the numerous challenges that this poses to automated assessment, existing methods commonly apply natural language processing (NLP) to build a high-dimensional representation of student responses that is then combined with various machine learning approaches (e.g. [13, 15, 3, 5]).

In this paper, we observe one such prior works on automated assessment model of student open responses in mathemat-

S. Baral. Improving automated assessment and feedback for student open-responses in mathematics. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 795–798, Durham, United Kingdom, July 2022. International Educational Data Mining Society.


https://doi.org/10.5281/zenodo.6853036

ics based on sentence-level semantic representation of the student open responses: "SBERT-Canberra" method. With the goal of exploring the limitations and further improving the method for assessment, we discuss our prior study that applies an exploratory error analysis to identify the areas of improvement that may be addressed by future iterations of these methods. We further propose a simple solution to improve upon the SBERT-Canberra method for automated assessment by addressing one of these error factors, that is the presence of mathematical terms and expressions in student answers. Additionally we seek to explore and address other factors of error to further improve the SBERT-Canberra method for auto-scoring. Finally we also intend to improve and expand this work towards feedback recommendations for student open-responses in mathematics.

## 2. PRIOR WORK

For this research, we observe one of the prior works on automated assessment of student open-responses in mathematics: the "SBERT-Canberra" model . This method follows a simple similarity-ranking procedure to generate the score predictions based on Sentence-BERT (SBERT) [12]. When suggesting a score for a given student response, it first applies SBERT to generate a high-dimensional feature embedding that describes the response as a whole. The intuition behind this is to capture semantic and syntactic meaning within this embedding, such that similar responses would be mapped closer within the embedding space. The SBERT embedding for this new student response is then compared to SBERT embeddings corresponding to a pool of historic labeled student responses utilizing the canberra distance measure [9]. In the final step the score for the historic response corresponding to the smallest distance (i.e. the most similar response) is used as the score prediction. The intuition behind this method is that similar answers to the a problem would have the same score.

## 3. CURRENT WORKS

### 3.1 Error Analysis of Auto-Scoring Method

With the goal of exploring the limitations of the SBERT-Canberra approach in order to identify the areas where the model does well and where it may yet improve through future iteration, we conducted an exploratory error analysis of the method. The dataset for the analysis was collected during the pilot testing of a teacher-augmentation tool designed to aid the assessment of open-responses within the ASSISTments[7] online learning platform. This tool, called QUICK-Comments used the SBERT-Canberra model to predict the scores for student open-responses in mathematics. Toward the error analysis, we observe two regression models that observe absolute model error as a dependent variable. The absolute model error here is the absolute difference between the score predicted by the SBERT-Canberra model and the teacher assigned grade for a particular student answer.

#### *3.1.1 Uni-level linear model*

For the uni-level model we explore characteristics of student answers in the context of this modeling error. These answer-level features are composed of length of answer, average character per word in the answer, total nos. of numbers, and operators in the answer text, percentage of mathematical expression in the answer text, and presence of images. The results of the error analysis are presented in Table 1. It is found that the uni-level linear model with student answer level features explains 38.6% of the variance of the outcome as given by r-squared. Out of the six student answer-level features, nearly all were found to be statistically reliable predictors of model error. However, only two of these variables: Equation Percent and Presence of Images were found to have more meaningful coefficient as compared to other features. This suggested that the presence of mathematical expressions and images(unsurprisingly) both correlate with higher prediction error.

#### *3.1.2 Multi-level linear model*

Similarly, we then apply a multi-level model to observe which of student-, problem-, and teacher-level identifiers most explains any observed modeling error. In regard to this, accounting for student, problem, and teacher identifiers each as random effects, we see that the inclusion of these level-2 factors explains some of the impact of the fixed effects (Table 1). It is worth noting that the level-2 variables account for 55.5% of the variance of the outcome. This suggests that a majority of the modeling error can be explained by the factors external to the student answers. Looking at the variance of the random effects, it can be seen that the problem level identifiers contribute most in terms of explaining the variance of the outcome.

### 3.2 Improving the Auto-Scoring method

With the results from the error analysis of the SBERT-Canberra method for auto-scoring, next we seek to improve this approach addressing for the factors that contributed to the modeling error. We know from the error analysis that one of the limitations of the SBERT-Canberra method in predicting the scores is in the presence of mathematical terms and equations in the student answer. To address this limitation, we propose the "Math Term Frequency" (MTF) model, drawing inspiration from assessment methods applied for close-ended problems. The goal of this method is to learn about the mathematical terms present in student answers and supplement this method to the previously developed SBERT-Canberra model through ensembling. For this, first we identify non-linguistic terms in students answers, and then identify the most frequently-occurring terms for each possible integer score to learn a kind of rubric. These most frequently occurring non-linguistic terms are then used to develop the features for this method. These features indicate whether a newly-observed student response contains any of the most frequent terms most commonly associated with each given score. They are finally used in a multinomial logistic regression (treating each score as an independent category), trained separately for each problem.

The score predictions from the MTF model are then ensembled with the SBERT-Canberra predictions using another logistic regression model, referred to as the SBERT-MTF model; to clarify, this ensemble regression model observes ten features corresponding to the probability estimates produced for each of the five possible scores for each of the two observed models. The goal of this is to combine the semantic representation captured by the SBERT method, while taking advantage of the non-linguistic term matching from the MTF method.

**Table 1: The resulting model coefficients for the uni-level linear regression model and random and fixed effects of the multi-level linear model of absolute error for auto-scoring method.**

| | Uni-level Linear | | Multi-level Linear | |
|---|---|---|---|---|
| | Variance | Std. Dev. | Variance | Std. Dev. |
| *Random Effects* | | | | |
| Student | — | — | 0.034 | 0.185 |
| Problem | — | — | 0.313 | 0.559 |
| Teacher | — | — | 0.048 | 0.851 |
| | B | Std. Error | B | Std. Error |
| *Fixed Effects* | | | | |
| Intercept | 0.581*** | 0.017 | 0.772*** | 0.070 |
| Answer Length | -0.008*** | 0.001 | -0.009*** | 0.001 |
| Avg. Word Length | -0.014*** | 0.003 | -0.013** | 0.003 |
| Numbers Count | <0.001 | <0.001 | <0.001 | <0.001 |
| Operators Count | -0.006*** | 0.001 | 0.002 | 0.001 |
| Equation Percent | 0.443*** | 0.018 | 0.080*** | 0.022 |
| Presence of Images | 2.248*** | 0.021 | 1.858*** | 0.028 |
| Answer Length X Images | -0.081*** | 0.004 | — | — |

*p <0.05 **p<0.01 ***p<0.001

# 4. FUTURE WORKS

In this research, we have identified areas where more advanced methods of image processing and natural language processing (or math language processing), may lead to further improvements in the existing methods for automated assessment. We have proposed a simple solution to address the limitations of current scoring method in presence of non-linguistic terms in student answers. While this proposed solution is an initial step towards addressing mathematical terms in these NLP based methods, we intend to explore more advanced methods based on Mathematical language processing and MathBERT to address such issues in future.

Further, we believe this method can be extended to recommend feedback messages in addition to suggesting numeric scores. With this, the next steps in our research is to expand the existing methods in suggesting and generating directed feedback to these student answers. We believe that the proposed MTF method combined with SBERT-Canberra can be extended as a prediction task in predicting whether given student responses are similar or not. This could be further beneficial in finding similar answers to math open-ended questions and thus utilizing this in improving the feedback recommendation task.

While the current works are based on textual open-ended responses, there are other forms of open-ended responses in mathematics including drawn diagrams and graphs, hand written formulas and expression uploaded as images, and other forms of audio and video responses. We seek to expand this research to these other forms of student open-ended responses, and further study the feasibility in deploying these automated methods in a computer-based learning environment.

# 5. ACKNOWLEDGMENTS

We thank multiple NSF grants (e.g., 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, 1535428, 1440753, 1316736, 1252297, 1109483, & DRL-1031398), as well as the US Department of Education for three different funding lines; the Institute for Education Sciences (e.g., IES R305A170137, R305A170243, R305A180401, R305A120125, R305A180401, & R305C100024), the Graduate Assistance in Areas of National Need program (e.g., P200A180088 & P200A150306 ), and the EIR. We also thank the Office of Naval Research (N00014-18-1-2768) and finally Schmidt Futures we well as a second anonymous philanthropy.

# 6. REFERENCES

[1] J. Burstein, C. Leacock, and R. Swartz. Automated evaluation of essays and short answers. 2001.

[2] E. Chen, M. Heritage, and J. Lee. Identifying and monitoring students' learning needs with technology. *Journal of Education for Students Placed at Risk*, 10(3):309–332, 2005.

[3] H. Chen and B. He. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752, 2013.

[4] L. Cutrone, M. Chang, et al. Auto-assessor: computerized assessment system for marking student's short-answers automatically. In *2011 IEEE International Conference on Technology for Education*, pages 81–88. IEEE, 2011.

[5] S. Dikli. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1), 2006.

[6] J. A. Erickson, A. F. Botelho, S. McAteer, A. Varatharaj, and N. T. Heffernan. The automated grading of student open responses in mathematics. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 615–624, 2020.

[7] N. T. Heffernan and C. L. Heffernan. The assistments ecosystem: Building a platform that brings scientists

and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, 2014.

[8] K. Holstein, G. Hong, M. Tegene, B. M. McLaren, and V. Aleven. The classroom as a dashboard: Co-designing wearable cognitive augmentation for k-12 teachers. In *Proceedings of the 8th international conference on learning Analytics and knowledge*, pages 79–88, 2018.

[9] G. Jurman, S. Riccadonna, R. Visintainer, and C. Furlanello. Canberra distance on ranked lists. In *Proceedings of advances in ranking NIPS 09 workshop*, pages 22–27. Citeseer, 2009.

[10] A. S. Lan, D. Vats, A. E. Waters, and R. G. Baraniuk. Mathematical language processing: Automatic grading and feedback for open response mathematical questions. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 167–176, 2015.

[11] M. Mohler and R. Mihalcea. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 567–575, 2009.

[12] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[13] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. Lee. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–168, 2017.

[14] L. Zhang, Y. Huang, X. Yang, S. Yu, and F. Zhuang. An automatic short-answer grading model for semi-open-ended questions. *Interactive learning environments*, 30(1):177–190, 2022.

[15] S. Zhao, Y. Zhang, X. Xiong, A. Botelho, and N. Heffernan. A memory-augmented neural model for automated grading. In *Proceedings of the fourth (2017) ACM conference on learning@ scale*, pages 189–192, 2017.