Enhancing Auto-scoring of Student Open Responses in the Presence of Mathematical Terms and Expressions

Sami Baral[1]*, Karthik Seetharaman[1]*, Anthony F. Botelho[2], Anzhuo Wang[1], George Heineman[1], and Neil T. Heffernan[1]

[1] Worcester Polytechnic Institute, Worcester MA, USA {sbaral, kvseetharaman, awang6, heineman, nth}@wpi.edu

[2] University of Florida, Gainesville FL, USA {a.botelho}@ufl.edu

Abstract. Prior works have led to the development and application of automated assessment methods that leverage machine learning and natural language processing. The performance of these methods have often been reported as being positive, but other prior works have identified aspects on which they may be improved. Particularly in the context of mathematics, the presence of non-linguistic characters and expressions have been identified to contribute to observed model error. In this paper, we build upon this prior work by observing a developed automated assessment model for open-response questions in mathematics. We develop a new approach which we call the "Math Term Frequency" (MTF) model to address this issue caused by the presence of non-linguistic terms and ensemble it with the previously-developed assessment model. We observe that the inclusion of this approach notably improves model performance, and present an example of practice of how error analyses can be leveraged to address model limitations.

Keywords: Math-terms · Open-ended responses · Automated assessment · Machine Learning · Natural Language Processing · Mathematics.

1 Introduction

Advancements in artificial intelligence and machine learning research have led to greater integration of prediction models into educational contexts through computer-based learning systems. These systems are being used in educational settings to support teachers and students in a variety of ways. Most prominent of the supports offered by most learning systems is that of automated assessment.

When assessing open-ended problems, however, the correctness of student responses can be subjective, where teachers commonly assess students based on an explicit or implicit rubric that identifies key points that must be included in a student response to sufficiently demonstrate comprehension. Current automatic assessment methods commonly apply natural language processing (NLP)

* Both authors contributed equally to this research

to build a high-dimensional representation of student responses that is then combined with various machine learning approaches (e.g. [8, 11, 2, 3]).

In consideration of the challenges in assessing open-ended problems, mathematics based domains make developing automated assessment models even more difficult, as most traditional NLP techniques were not designed for such a context, with a few recent exceptions [4, 7, 9]. Recent work has identified that the existence of non-linguistic terms is positively correlated with model prediction error in models that have outperformed existing benchmarks in this context [1].

This work presents a simple, targeted method to resolve this problem. We call this proposed method the "Math Term Frequency" (MTF) model and demonstrate how it can be combined with previously-developed assessment models to improve performance. Specifically, this work addresses the following research questions: 1) How does accounting for non-linguistic terms through our MTF model affect the performance of auto-assessment methods on existing benchmarks? and 2)Does our MTF method reduce the correlation between non-lingustic terms and model prediction error?

2 Dataset

To explore and examine the methods proposed in this work, we observe two datasets consisting of student answers to mathematics open-response questions. These datasets were collected from ASSISTments [6] and contains 150,477 student responses from 27,199 students for 2,076 open-ended math problems scored by 970 unique teachers (where each response was scored by a single teacher); this dataset is the same used to establish benchmark results [4] and is used to directly compare performance against models presented in prior work[4, 1]. Teachers scored responses based on a 5-point integer scale ranging from 0 to 4 , with a 4 indicating a very strong and a 0 indicating a very weak response. The second dataset used in this paper was similarly used in prior work to conduct an error analysis to identify factors that correlate with prediction error [1]. This dataset is comprised of student open responses collected in a pilot study of the QUICK-Comments tool and contains 30,371 scored student responses from 1,628 students for 915 unique open-response questions assessed by 12 teachers.

3 The SBERT-MTF Model

The methods presented in this work target the specific problem of non-linguistic terms contributing to prediction error. The previously-developed SBERT-Canberra model outperformed previous decision-tree- and deep-learning-based approaches [4] by leveraging pre-trained Sentence-BERT embeddings. The challenge, however, is that only a finite number of words (and sentences, by extension) can be recognized by these methods. When observing non-lingustic terms such as numbers and expressions, many such terms may not be represented within the embeddings (e.g. representing "the answer is 4.3333" with the same embedding as, for example, "the answer is 2.987" if neither of the numbers are recognized).

Instead, we propose the "Math Term Frequency" (MTF) method which takes a much simpler approach, drawing inspiration from assessment methods applied for close-ended problems. The goal of this method is to supplement the previously-developed SBERT-Canberra model through ensembling, resulting in what we are calling the "SBERT-MTF" model.

The MTF method works by first parsing student answers to identify non-linguistic terms. The function[3] splits each student answer by spaces, removes alphabet-only terms (accounting for punctuation), removing spaces around math operators, and rounding off large decimals. Once the non-linguistic terms have been identified, the MTF method involves identifying the most frequently-occurring terms for each possible integer score as a means of learning a kind of rubric. There will likely be some terms that are common throughout all scored answers, but there are likely to be some terms that demonstrate comprehension; similarly, students exhibiting common misconceptions may arrive at a similar set of incorrect answers. With this in mind, we select the five most-frequent terms from the list of parsed non-linguistic terms for each problem. With these, for a new response for which we want to generate a score, we calculate a set of 5 indicator values representing whether the response contains each of the most-frequent terms. These features are used in a multinomial logistic regression (following previous works) that is 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.

3.1 SBERT-MTF Model Performance

As to directly compare the existing method to the prior works, we use similar evaluation method and dataset used in [1, 4]. This evaluation method utilizes a 2-parameter IRT model to compare model estimates [10]. The model predictions are used as covariates within the IRT model allowing for the comparison of scoring methods that controls for variables of general student ability and problem difficulty; the number of words in the response is also added as a covariate in this evaluation model in an attempt to further compare models on their ability to interpret student answers rather than be based on other more superficial response features. This evaluation method allows for a fair comparison that accounts for factors that likely impact score that are external to the observed text of the student response. For comparison to previous works, we evaluate our method using three metrics: AUC (see [5]), Root Mean Squared Error (RMSE; calculated using model estimates as a continuous-valued integer scale), and Cohen's Kappa.

[3] All code used in this work is available at *https://github.com/ASSISTments/SBERT-MTF*

The IRT model performance of the Math terms frequency model as compared to the performance of the prior models for scoring open-ended responses is presented in Table 1. The results suggests that the proposed SBERT-MTF model outperforms the previous highest-performing model across evaluation metrics.

Table 1. IRT Model Performance compared to the models developed in prior works related to auto-scoring of student open responses in mathematics.

Model	AUC	RMSE	Kappa
Baseline IRT	0.827	0.709	0.370
IRT + SBERT-Canberra	0.856	0.577	0.476
IRT + SBERT-MTF	**0.871**	**0.524**	**0.508**

3.2 Error Analysis of SBERT-MTF

The proposed MTF method was designed to address a very targeted problem exhibited by the previously-developed SBERT-Canberra model. We therefore conduct a similar error analysis to observe whether this method impacts the observed positive correlation between the presence of non-linguistic terms and model error. For this analysis, we use the second dataset as described in Section 2 for a direct comparison with the previous work. While the modeling task treats scoring as a categorization task, we convert the model predictions to a ordinal-scale integer value (i.e. 0-4). We calculate model prediction error as the absolute value of the teacher-provided score minus the predicted score. In this way, positive values correspond with higher error and values close to 0 represent low error (high performance) and conduct a linear regression observing absolute error as the dependent and answer-level features as independent variables.

We compare three models within this analysis to identify how two modeling decisions presented in this work correspond with observed changes in feature coefficients. The first model observed is that of the SBERT-Canberra model reported in [1] as a baseline for comparison. The second model uses the same SBERT-Canberra method, but trains a logistic regression per problem with the model predictions as covariates (e.g. similar to the ensembled method described earlier, without MTF); the intuition here is that problem-specific adjustments may itself help to account for error in the model. Finally, we observe the ensembled SBERT-MTF model for impacts beyond these other two methods.

The results of the error analysis is presented in Table 2. The results indicate that the linear model for both Logistic SBERT and SBERT-MTF explains 34.8% of the variance of the outcome as given by r-squared; this alone suggests that there is a large portion of variance in the error unexplained by the observed features. Among the observed features, similar to the results from [1], nearly all were statistically reliable in predicting the model error. However, it is arguable that from the relatively small scale of most coefficients, two of the features exhibit

Table 2. The resulting model coefficients for the uni-level linear regression model of absolute error for SBERT Canberra, Logistic SBERT and MTF model.

	SBERT-Canberra		Logistic SBERT		SBERT-MTF	
	B	Std. Error	B	Std. Error	B	Std. Error
Intercept	0.581***	0.017	0.738***	0.017	0.776***	0.070
Answer Length	-0.008***	0.001	-0.008***	0.001	-0.009***	0.001
Avg. Word Length	-0.014***	0.003	-0.013***	0.003	-0.014***	0.003
Numbers Count	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Operators Count	-0.006***	0.001	0.001	0.001	0.004**	0.001
Equation Percent	0.443***	0.018	-0.062***	0.019	-0.128***	0.019
Presence of Images	2.248***	0.021	2.058***	0.022	2.018***	0.022

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

more meaningful impacts in comparison to the others: the presence of mathematical expression and presence of images in the student answers. However, with the introduction of a logistic regression model that follows the SBERT-Canberra method, the coefficient value of presence of mathematical terms has changed; it would appear that accounting for problem-level adjustments alone removes much of the impact of non-linguistic terms in the dataset. Most notably, however, is that the addition of our MTF method exhibits an even stronger negative correlation between the presence of non-linguistic terms and model error; what once was a weakness now appears to be a potential strength of the model.

4 Discussion and Future Work

The results of all of the presented analyses illustrate MTF (specifically, SBERT-MTF) as a promising method to mitigate model error attributed to the presence of non-linguistic terms. The MTF method represents an intentionally-simple approach to address a targeted weakness observed in previously-developed models and seemingly led to positive impacts.

With that, there are still several areas in which these models could be improved, in addition to improving the accuracy of the parsing function. Most notably, is the remaining correlation between the presence of images and model error. While this is not surprising, as the models do nothing to account for images, this remains an unhandled case that cannot be ignored. As it is also the case that some students include mixtures of natural language, non-linguistic terms, and images all in the same answer, developing methods to handle such cases fairly is important for future work.

Similarly, the error analysis suggests that there is a large amount of variance in model error left unexplained. Previous work [1] identified problem- and teacher-level factors that seemingly account for much of this unexplained error, but this does not provide clear guidance as to how to account for these external factors fairly within an automatic assessment model.

5 Acknowledgements

We thank multiple grants (e.g., 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, 1535428, 1440753, 1316736, 1252297, 1109483, & DRL-1031398); IES R305A170137, R305A170243, R305A180401, R305A120125, R305A180401, & R305C100024, P200A180088 & P200A150306, as well as N00014-18-1-2768, Schmidt Futures and a second anonymous philanthropy.

References

1. Sami Baral, Anthony Botelho, John Erickson, Priyanka Benachamardi, and Neil Heffernan. Improving automated scoring of student open responses in mathematics. In *Proceedings of the Fourteenth International Conference on Educational Data Mining, Paris, France*, 2021.
2. Hongbo Chen and Ben 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.
3. Semire Dikli. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1), 2006.
4. John A Erickson, Anthony F Botelho, Steven McAteer, Ashvini Varatharaj, and Neil 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.
5. David J Hand and Robert J Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45(2):171–186, 2001.
6. Neil T Heffernan and Cristina Lindquist 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.
7. Andrew S Lan, Divyanshu Vats, Andrew E Waters, and Richard 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.
8. Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chungmin 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.
9. Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Ben Graff, and Dongwon Lee. Mathbert: A pre-trained language model for general nlp tasks in mathematics education. *arXiv preprint arXiv:2106.07340*, 2021.
10. Wijbrandt van Schuur. *Ordinal Item Response Theory: Mokken Scale Analysis*, volume 169. SAGE Publications, 2011.
11. Siyuan Zhao, Yaqiong Zhang, Xiaolu Xiong, Anthony Botelho, and Neil Heffernan. A memory-augmented neural model for automated grading. In *Proceedings of the fourth (2017) ACM conference on learning@ scale*, pages 189–192, 2017.