Natural Language Processing for Open Ended Questions in Mathematics within Intelligent Tutoring Systems

John A. Erickson
Worcester Polytechnic Institute
100 Institute Road
Worcester MA, 01609
jaerickson@wpi.edu

ABSTRACT

Intelligent tutoring systems continue to enable teachers insight into their students in an immediate fashion. With deep fine-grained data provided to the teachers, they can gain a deeper understanding of the student's learning. While multiple systems exist, most are limited to specific, close-ended questions; these include questions with a set of known acceptable answers, such as solving for 'x' in an equation (i.e. in 'x+4=6', the clear answer would be 2). Questions of this variety are implemented within these systems and allot for timely feedback to the students. A system can easily decipher certain values to be incorrect answers and help can be offered to the student. While close-ended problems provide a wide range of insights into the student's process, they are often unable to gain the deeper discernment of the student's understanding. Open response questions elicit a greater scope of the student's understanding. However, very few intelligent tutoring systems provide support to teachers and students for these types of questions. Within the few that can, they are not able to offer automation for the process. One of the greater appeals of computer-based systems is that they provide teachers automated grading and give students immediate feedback. It is therefore my goal to further the study and development of automated assessment and feedback tools to support open-ended problems within computer-based systems. Toward this goal, my focus of research is on the development and deployment of automatic grading models, exploration of fairness within such models, and expansion of existing systems to leverage this research.

Keywords

Natural language processing; machine learning; word-embeddings; intelligent tutoring systems; automated grading; automated feedback

1. INTRODUCTION

Intelligent tutoring systems (ITS) have been around for some time, and their benefits have been discussed and noted in studies such as [13][17]. These benefits, however, have been limited to close-ended problem types. As such, problems with close-ended answers are at the core of most ITS; including ASSISTments [5], McGraw Hill's ALEKS^{TM} and Carnegie Learning's Cognitive Tutor^{TM}. This limitation comes from the overall goal of ITS; to provide automated feedback to students and timely reports to teachers about their students. Questions with close-ended answers allow these systems to achieve this goal. For instance, its very simple to set up a system to understand the correct answers when 1/2 or .5 are the only acceptable student answer. Studies such as [14] have discussed why multiple choice questions (close-ended questions) are so appealing: they're easy, accurate and timely to grade. While it is evident that the teachers gains a substantial understanding of the students comprehension from these questions, there is more to student's process of thinking. If the student selects A, the teacher can assert the student's rationale; however, this is a summation from other students selecting the same answer. Open responses questions provide students the opportunity to explain their own personal rationale; giving teachers an even more in depth understanding of the student's process of thinking. Studies such as [6] called attention to the fact that there are vast advantages to a greater spectrum of questions types; when focusing on evaluations with a single question type, it's insufficient in testing the students actual understanding and rationale/critical thinking. By providing support for open response questions, teachers are able to discern, in greater detail, what point the student became confused or if they ever understood. This is also supported by [7] which discussed the wider range of cognition required with open response questions as compared to close-ended multiple choice questions. However, as mentioned earlier, few intelligent tutoring systems support this type of question.

While not the only system to support open response problems, ASSISTments, the system through which much of my prior research has been conducted, is developing tools to improve the support of these problems for teachers. The capability to automatically grade student answers or provide immediate feedback to students is still lacking in comparison to what is possible for close-ended problems. For open response questions, natural language processing (NLP) must to be utilized to provide such tool. Additionally, the infrastructure needs to be in place to support these machine learning algorithms for real-time use within classrooms.

John Erickson "Natural Language Processing for Open Ended Questions in Mathematics within Intelligent Tutoring Systems" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 762 - 765

Table 1: Rasch Model Performance from Erickson et al., 2020

Model	AUC	RMSE	Kappa
Rasch Model with teacher component	0.696	1.09	0.162
Rasch Model without covariates	0.827	0.709	0.370
Rasch Model with number words covariates	0.829	0.696	0.382
Rasch Model number words and Random Forest covariates	**0.850**	**0.615**	**0.430**
Rasch Model number words and XGBoost covariates	0.832	0.679	0.390
Rasch Model number words and LSTM covariates	0.841	0.637	0.415

In this paper I will be discussing my previous work, which has attempted to develop machine learned models for automatically grading student open response questions within ASSISTments, in addition to current and proposed future projects pertaining to the further study and development of tools to support these problems in classroom settings. Among this proposed future research, I will describe my intention to study similarity measures to allow students to see similar open ended response rationales to theirs; in this regard, I have drawn inspiration from an existing system, known as myDALITE [2], and propose an extension of this idea utilizing open-ended response problems.

2. PREVIOUS CONTRIBUTIONS

It's clear there is an advantage to developing a tool which can assist in automating open responses in mathematics within intelligent tutoring systems (ITS). If we can bridge the gap between the ITS capabilities within close-ended problems and open response problems, we can further empower teachers with a deeper knowledge of their student's logic. With this, I focused on starting from the ground up. Exploring our ability, within ASSISTments, to automatically grade open response student answers within open response questions in a mathematical domain.

2.1 Automated Grading

While others have utilized a multitude of NLP approaches to interpret and grade open response questions, [16] [15] [12] [18] , most have been working with non-mathematical content. Much of the NLP research has consisted of essays and sentences with a standard corpus. This is why so many approaches looked to utilizing deep learning approaches, such as word embeddings Word2Vec [8] and GloVe [10] to gain a vector relational understanding of words. For my research [3], I set out to automatically grade open response questions within the mathematical domain. Contrary to previous research, the corpus within this study was unique in the sense that student answers would be a diverse assortment of words and mathematical functions. Not only was the corpus diverse in words and functions, but the answers were diverse in length. Some student answers consisted of one or two words, while others responded with multiple sentences.

Within this research, the route was taken to approach the NLP task with a wide variety of approaches and methods. With models developed from traditional NLP approaches such as a term-frequency inverse document frequency, *tf-idf* (bag of words model which counts the number of occurrences of the word and re weights the word), to deep learning approaches with word embeddings, a wide spectrum of approaches were attempted.

Overall, 6 different models were developed to predict the student's grade on an open response mathematics question. In Table 1, the baseline model was a Rasch model which didn't take into account any NLP developed models. From there, we supplemented the Rasch model with a *teacher component* and *number of words* covariate. Each of those performed worse than either the tf-idf, or the word embedding approaches. By augmenting the models with NLP approaches, the Rasch model was able to improve and provide a stronger performing model with our data (c.f. [3] for further detail pertaining to this study and analyses).

3. CURRENT WORK

While the top performing model in my previous study showed promise with an AUC of .850, RMSE of 0.615, and Kappa 0.430, beating the baseline and all other models, it was decided to ensemble the 3 top performing models. The ensemble, along with the individual previous 3 top models, are now currently being used within a randomized control trial and integrated within ASSISTments. What has become more and more evident is that when utilizing pre-trained word embeddings, there needs to be close consideration of model fairness. As studies such as [1] noted, there can be underlying biases within word embedding models.

3.1 Assessing Fairness

Since multiple of the models within the automated grading study utilize pre-trained word embeddings, my research has progressed towards exploring potential bias within our models. Its imperative that models being implemented within an ITS, or any study, should minimize bias; especially as it pertains to grading. The grades should be based solely on the content, nothing else. As stated earlier, [1] notes that it doesn't matter which embedding approach you use (or pre-trained embeddings in our case), biases, such as gender bias, can sneak in. As the paper references, embeddings can teach models that woman is to homemaker as male is to computer programmer. This is something we explicitly want to avoid in any predictive models within an ITS.

Currently, work is being done to identify potential bias within models from my previous automated grading study. What is imperative is to be able to clearly identify the bias, if there is evidence of bias, and if its coming from pre-trained word embedding (when we account for the different word usages of males and females) or the models the grade predictions are trained with. By developing steps to directly compare models, and word representations, to predict grades given women responses/male responses, we can hopefully identify whether bias is present. We are building our approach from prior works (c.f. [4][9]), and if we can clearly identify which

models have the least amount of bias, then we push those models to production. Additionally, we will be exploring how to handle the bias, if needed, within the suspected models.

3.2 Randomized Control Trial

Currently, my research also is simultaneously being applied to a randomized control trial. This is a study in which the automated grading models are being used to provide student's with their potential grade before they submit an answer. So, once the student's have submitted a answer to the open response mathematics question, one of the conditions will take the strongest performing grade prediction model for the problem and suggest a grade. This grade is then presented to the student and the student will be presented with the option to edit their answer. This poses many interesting questions such as: will the student's edit their answers? If they do, by how much have the answers changed and how much has their grade changed. This is ongoing research and I will continue to develop new models and take into consideration the bias study previously discussed here.

3.3 Comment Suggestions

As discussed previously, one of the main attractions to ITS is the automation. While I have presented multiple models that predict the students grade with reasonable accuracy, within our data, it is clear there is another step. Providing automated feedback is the next optimal tool for teachers and students. Currently, work has been done developing an approach which suggest responses by utilizing similarity calculations. Recently, our team collected data where teachers graded a set of student open response answers. This allowed us to have multiple teachers grade the same student answer, as well. Within this, teachers would grade and create a category which they would place the student answer in. This was performed across multiple problems.

With this, there is now a more robust dataset of answers and associated teacher responses. By utilizing similarity calculations, ranging from Levenshtein distances to SBERT [11], when a student submits an answer to a problems (one which we have previous data on) the most similar student answer on file is calculated and we then can suggest those associated teacher responses with that most similar answer.

Additionally, its being explored how these methods could be validated. For instance, aside from manually looking at the suggested responses, how could there be an offline evaluation of these methods (that does not require teachers to select from the undoubtedly poor suggestions produced by early iterations of such a tool). For each problem, the 3 most similar answer for each individual answer (which has been graded and categorized by our teachers) are selected using both SBERT and Levenshtein distances. From there, it is calculated how many of the teacher categories are the same for the similar answer and the original answer. The method with the most agreement, for each problem, is selected to use for future student answers for said problem.

4. FUTURE WORK

With accurate grade prediction models, a potential method to identify bias, and an approach to selecting similar student answers, I have a set of approaches which lends itself to the next step I wish to take. I am looking to explore whether we can expand upon just suggesting the student to go back and edit (the randomized controlled trial); can we use NLP to take the students answer, discover which are the most similar, find those similar answers and share their rationale with the student. Then allowing the students the opportunity to go back and either chose their submission or re-write their answers to reflect what they have learned from other similar (or possibly dissimilar) answer rationale. This requires a similarity calculation, a grade predictions (to see if the student's answers and most similar answer would retain the same or different grade) and then a way to show are calculations are accurate. Then once the student's answer has a top 3 similar student answers, the rationale (not answers) are shared. As identified earlier, this practice is in-part analogous to how an existing system, myDALITE [2] functions. It is for this reason that these same methods might be suited to expand upon this idea to provide teachers with new tools that can be used in the classroom.

In this system, students are presented with a multiple choice question and asked to provide an explanation, or rationale, for their work. Students are then presented with other rationales and asked if they would like to keep their answer or if a rationale for a different response has convinced them to change their answer. I wish to explore if this approach could be performed with open response questions. Instead of an initial multiple choice question the student writes a answer and rationale to an open response question and then similar responses are presented, giving the student the option to either change their response or continue. This would require multiple of my previous and current research to prepare such a approach.

This would be a fascinating exploration into how confident a student is in their response. If after seeing others rational, does that convince students to re-evaluate or edit their answers? We may be able to explore what types of answers are confident answers and how much they differ from less confident answers. Additionally, I would like to continue to use NLP to help identify gaming behavior with this type of system; it would be important to identify students answering with "I don't know" types of responses and avoid them simply being presented with other rationales. There are also questions into whether seeing other's rationale could hurt the students learning and cause more confusion. This is an aspect of the study which would need to be expanded upon.

Overall, there have been direct effects of my research, including the implementation of the automatic grader in ASSISTments using the models built in my previous research. Additionally, the current RCT provides an opportunity to see how these predicted grades could impact a student's answer if they were exposed to the grade. Lastly, there is potential for my work calculating similarities between student answers to impact how ASSISTments suggest responses for teachers to students. Hopefully, saving the teacher time and increasing the amount of open response questions given out.

5. ACKNOWLEDGEMENTS

I thank multiple NSF grants (e.g., 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, 1535428, 1440753, 1316736, 1252297, 1109483, & DRL-1031398), the US Department of Education Institute for Education Sciences (e.g., IES R305A170137, R305A170243, R305A180401, R305A120125, R305A180401, & R305C100024) and the Graduate Assistance in Areas of National Need program (e.g., P200A180088 & P200A150306), and EIR the Office of Naval Research (N00014-18-1-2768 and other from ONR) and finally Schmidt Futures.

6. REFERENCES

[1] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[2] E. S. Charles, N. Lasry, S. Bhatnagar, R. Adams, K. Lenton, Y. Brouillette, M. Dugdale, C. Whittaker, and P. Jackson. Harnessing peer instruction in-and out-of class with mydalite. In *Education and Training in Optics and Photonics*, page 11143_89. Optical Society of America, 2019.

[3] 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.

[4] J. Gardner, C. Brooks, and R. Baker. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 225–234, 2019.

[5] 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.

[6] K. Y. Ku. Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking skills and creativity*, 4(1):70–76, 2009.

[7] M. E. Martinez. Cognition and the question of test item format. *Educational Psychologist*, 34(4):207–218, 1999.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[9] J. Ocumpaugh, R. Baker, S. Gowda, N. Heffernan, and C. Heffernan. Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, 45(3):487–501, 2014.

[10] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[11] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[12] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. M. 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.

[13] J. Roschelle, M. Feng, R. F. Murphy, and C. A. Mason. Online mathematics homework increases student achievement. *AERA Open*, 2(4):2332858416673968, 2016.

[14] M. G. Simkin and W. L. Kuechler. Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, 3(1):73–98, 2005.

[15] J. Z. Sukkarieh and J. Blackmore. c-rater: Automatic content scoring for short constructed responses. In *Twenty-Second International FLAIRS Conference*, 2009.

[16] J. Z. Sukkarieh, S. G. Pulman, and N. Raikes. Automarking: using computational linguistics to score short, free text responses. 2003.

[17] K. VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.

[18] 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. ACM, 2017.