Exploring Automatic Short Answer Grading as a Tool to Assist in Human Rating

Aubrey Condor^(✉)

University of California, Berkeley, Berkeley, USA
aubrey_condor@berkeley.edu

Abstract. This project proposes using BERT (Bidirectional Encoder Representations from Transformers) as a tool to *assist* educators with automated short answer grading (ASAG) as opposed to replacing human judgement in high-stakes scenarios. Many educators are hesitant to give authority to an automated system, especially in assessment tasks such as grading constructed response items. However, evaluating free-response text can be time and labor costly for one rater, let alone multiple raters. In addition, some degree of inconsistency exists within and between raters for assessing a given task. Recent advances in Natural Language Processing have resulted in subsequent improvements for technologies that rely on artificial intelligence and human language. New, state-of-the-art models such as BERT, an open source, pre-trained language model, have decreased the amount of training data needed for specific tasks and in turn, have reduced the amount of human annotation necessary for producing a high-quality classification model. After training BERT on expert ratings of constructed responses, we use subsequent automated grading to calculate Cohen's Kappa as a measure of inter-rater reliability between the automated system and the human rater. For practical application, when the inter-rater reliability metric is unsatisfactory, we suggest that the human rater(s) use the automated model to call attention to ratings where a second opinion might be needed to confirm the rater's correctness and consistency of judgement.

Keywords: Assessment · Automated short answer grading · BERT · Natural Language Processing · Inter-rater reliability

1 Introduction

Although it has been shown that incorporating constructed response items in educational assessments is beneficial for student learning [2], the burden of time spent grading constructed response activities, as opposed to that of multiple choice questions, can deter educators from their use. In addition, the quality of human ratings of student responses can vary in consistency and reliability [15]. Using an automated system for grading free-text could help to alleviate this time

burden as well as produce more consistent ratings. However, from the educator’s perspective, completely removing human judgement from assessment tasks is neither responsible nor realistic. Natural Language Understanding (NLU) models are not yet able to discern all the nuances of language as well as a human. In high-stakes grading situations, incorrect ratings can have dire consequences for students.

Recent Automated Short Answer Grading (ASAG) Research using the most state-of-the-art language models, trained on large quantities of data, is only able to predict human ratings correctly less than 85% of the time. Notable recent work includes Crossley et al. who used Latent Semantic Analysis (LSA) to assess student summarizations [3]. Mieskes et al. combined several different automated graders to create a superior ensemble grader [8]. Qi et al. created a hierarchical word-sentence model using a CNN and BLSTM model [9]. Sung et al. examined the effectiveness of pre-training BERT as a function of the size of training data, number of epochs and generalizability across domains [12]. In a separate study, Sung et al. pre-trained BERT on relevant domain texts to enhance the existing model for ASAG [11]. Dhamecha et al. introduced an iterative data collection and grading approach for analyzing student answers [5]. Finally, Hu et al. incorporated a technique called Recognizing Textual Entailment to investigate whether a given passage and question support the predicted answer [14].

We propose using a compressed version of the BERT model called bert-base to simplify the training process and show that with a relatively small amount of training data (less than 70 student answers per question), we can achieve high enough inter-rater reliability to assist a human grader in constructed response rating tasks.

2 Methods

2.1 Dataset

A data set called DT-Grade was used, consisting of short constructed answers from tutorial dialogues between students and an Intelligent Tutoring System called Deep Tutor, created at the University of Memphis Institute for Intelligent Systems [10]. About 1100 student responses, in 100 words or less, to conceptual questions relating to Newtonian Physics were randomly selected from 40, junior-level college students. Included in the data are 34 distinct questions with relative question context information. Initial ratings were completed by experts and each answer was annotated for correctness by categorizing it as one of four categories: correct, correct-but-incomplete, contradictory, and incorrect [1] (Fig. 1).

Question Context	Question	Student Answer	Rating
A car windshield collides with a mosquito ...	How does Newton's third law ...	the windshield will apply a force ...	Incorrect
A rocket pushes a meteor with constant ...	Can you articulate Newton's ...	if there is zero net force on ...	Correct
...	...	...	...
...	...	...	...

Fig. 1. A snippet of the DT-Grade dataset

2.2 Data Pre-processing

We removed all records from the dataset where the number of answers per question was less than 20. Remaining were 28 distinct conceptual physics questions, and the number of student responses per question ranged from 20 to 69. The filtered dataset then contained 994 records. We collapsed the four rating categories into two so that we will have a binary response variable in order to start with the simplest version of the model. Correct responses were considered correct, and all others (correct-but-incomplete, contradictory and incorrect) were considered incorrect. The question context text was concatenated with the question text as well as the student's answer text before creating the input vector embeddings. The concatenated input texts were tokenized using the bert-base-uncased tokenizer [16] (Fig. 2).

Fig. 2. Flow chart of data pre-processing

Training, validating, and testing data sets were created such that 70% of responses were randomly allocated to the training set, 15% to the validation set and 15% to the test set.

2.3 BERT Model

The language model we used, BERT, which stands for Bidirectional Encoder Representations from Transformers, was introduced in [4] as a revolutionary language representation model. It was the first to successfully learn by pre-training unlabeled text bidirectionally. Consequently, the model can be fine-tuned for many different tasks, such as ASAG, by adding only one additional layer to the existing deep neural network. We use a compressed version of the original BERT model, called bert-base, through a python package called fast-bert [13]. Fast-bert enables quick and simplified fine-tuning of the bert-base model for the assessment task at hand.

A simple grid search was used to tune the parameters and hyper-parameters of the model such that we achieve a high validation accuracy. The best results were observed with using a batch size of 8 (on a single GPU), a maximum

sequence length of 512, 8 training epochs and a learning rate of $6e-5$. In addition, the LAMB optimizer was used for training. Our particular ASAG task is essentially one of binary text classification - each response is classified as either Correct or Incorrect. Our model returns the predicted classification of response rating per input vector.

2.4 Cohen’s Kappa

We used Cohen’s Kappa as our metric for inter-rater reliability. It calculates the extent to which raters agree on rating assignments beyond what is expected by chance [7]. Cohen’s Kappa is calculated as follows:

$$k = \frac{p_0 - p_e}{1 - p_e} \quad (1)$$

where p_0 = the relative observed agreement among raters and p_e = the hypothetical probability of chance agreement. A k value of 0 represents agreement equivalent to random chance, and a value of 1 represents perfect agreement between raters.

3 Results

The best model achieved a testing accuracy of 0.760 and a Cohen’s Kappa statistic of 0.684. This represents the probability that the BERT model agrees with the human rater beyond random chance. With such a small amount of training data per question, we believe that these results provide evidence that transfer learning models such as BERT can remove a significant amount of human rating work, as well help achieve more consistent human ratings.

We must consider the question of whether an instructor would find the described system practical, and correspondingly whether the resulting Kappa statistic is good enough for real world use. One perspective is that, the human rater can incorporate context specific judgement about the extent to which they would like to examine the highlighted cases of disagreement. For example, if the assessment is used for low-stakes, formative purposes, it might not be practical for an educator to investigate rating mismatches in depth. However, if the questions will be used repeatedly in future assessments or the scoring is involved in a pass-fail discernment for a student, a detailed look into discrepancies may be appropriate.

4 Conclusion

In order for the field of education to adopt a willingness to embrace applicable research in Artificial Intelligence, researchers must consider the practicality and usefulness of new technologies from the educator’s perspective. Such technologies should act as a support for teachers; not as independent, decision-making entities. This project represents a work-in-progress to continually investigate how we can leverage artificial intelligence to be in service of human decision making.

References

1. Banjade, R., Maharjan, N., Niraula, N., Gautam, D., Samei, B., Rus, V.: Evaluation dataset (DT-Grade) and word weighting approach towards constructed short answers assessment in tutorial dialogue context. In: Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications, San Diego, CA, pp. 182–187. Association for Computational Linguistics (2016). <https://doi.org/10.18653/v1/W16-0520>
2. Chi, M.T., De Leeuw, N., Chiu, M.H., LaVancher, C.: Eliciting self-explanations improves understanding. *Cogn. Sci.* **18**(3), 439–477 (1994)
3. Crossley, S.A., Kim, M., Allen, L., McNamara, D.: Automated Summarization Evaluation (ASE) using natural language processing tools. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11625, pp. 84–95. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23204-7_8
4. Devlin, J., Chang, M., Le, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding (2018). [arXiv:hep-th/9910001](https://arxiv.org/abs/1810.03815)
5. Dhamecha, T.I., Marvaniya, S., Saha, S., Sindhgatta, R., Sengupta, B.: Balancing Human efforts and performance of student response analyzer in dialog-based tutors. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10947, pp. 70–85. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93843-1_6
6. Holmes, W., Bialik, M., Fadel, C.: Artificial intelligence in education: promises and implications for teaching and learning. Center for Curriculum Redesign, Boston, MA (2019)
7. Louis, M., Field, R.: Interrater agreement measures: comments on Kappan, Cohen’s Kappa, Scott’s π , and Aickin’s α . *Underst. Stat.* **2**(3), 205–219 (2003). <https://doi.org/10.1207/S15328031US0203>
8. Mieskes, M., Padó, U.: Work smart-reducing effort in short-answer grading. In: Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning, pp. 4294–4295. Linköping University Electronic Press, SLTC, Stockholm (2018)
9. Qi, H., Wang, Y., Dai, J., Li, J., Di, X.: Attention-based hybrid model for automatic short answer scoring. In: Song, H., Jiang, D. (eds.) SIMUtools 2019. LNICST, vol. 295, pp. 385–394. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32216-8_37
10. Rus, V., Niraula, N., Banjade, R.: DeepTutor: an effective, online intelligent tutoring system that promotes deep learning. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, pp. 4294–4295. AAAI Press (2015). <https://doi.org/10.5555/2888116.2888373>
11. Sung, C., Dhamecha, T., Saha, S., Tengfei, M., Reddy, V., Rishi, A.: Pre-training BERT on domain resources for short answer grading. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, CH, pp. 6071–6075. Association for Computational Linguistics (2019). <https://doi.org/10.18653/v1/D19-1628>
12. Sung, C., Dhamecha, T.I., Mukhi, N.: Improving short answer grading using transformer-based pre-training. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11625, pp. 469–481. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23204-7_39
13. Trivedi, K.: Fast-BERT. <https://github.com/kaushaltrivedi/fast-bert>. Accessed 25 Feb 2020

14. Hu, M., Wei, F., Peng, Y., Huang, Z., Yang, N., Li, D.: Read+ verify: machine reading comprehension with unanswerable questions. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6529–6537, July 2019
15. Wind, S., Peterson, M.: A systematic review of methods for evaluating rating quality in language assessment. *Lang. Test.* **35**(2), 161–192 (2018). <https://doi.org/10.1177/0265532216686999>
16. Wolf, T., et al.: HuggingFace’s transformers: state-of-the-art natural language processing. arXiv (2019) [arXiv:1910.03771](https://arxiv.org/abs/1910.03771)