Automated Short-Answer Grading Using Deep Neural Networks and Item Response Theory

Chapter · June 2020

DOI: 10.1007/978-3-030-52240-7_61

2 authors, including:

Masaki Uto
The Univ. of Electro-Communications
47 PUBLICATIONS **522** CITATIONS

SEE PROFILE

Automated Short-answer Grading Using Deep Neural Networks and Item Response Theory

Masaki Uto[0000−0002−9330−5158], Yuto Uchida

The University of Electro-Communications, Tokyo, Japan
uto@ai.lab.uec.ac.jp

Abstract. Automated short-answer grading (ASAG) methods using deep neural networks (DNN) have achieved state-of-the-art accuracy. However, further improvement is required for high-stakes and large-scale examinations because even a small scoring error will affect many test-takers. To improve scoring accuracy, we propose a new ASAG method that combines a conventional DNN-ASAG model and an item response theory (IRT) model. Our method uses an IRT model to estimate the test-taker's ability from his/her true-false responses to objective questions that are offered with a target short-answer question in the same test. Then, the target short-answer score is predicted by jointly using the ability value and a distributed short-answer representation, which is obtained from an intermediate layer of a DNN-ASAG model.

Keywords: Deep neural networks · item response theory · automated short answer grading.

1 Introduction

Short-answer questions are widely used to evaluate the higher abilities of test-takers, such as logical thinking and expressive ability. World-wide large-scale tests, such as the Test of English as a Foreign Language and the Graduate Management Admission Test, incorporate short-answer questions. However, the introduction of this type of question to these large-scale tests has prompted concerns related to scoring accuracy, time complexity, and monetary cost. Automated short-answer grading (ASAG) methods have attracted much attention as a way to alleviate these concerns [1, 2].

Conventional ASAG methods have relied on manually tuned features, which are laborious to develop [3, 10–12]. However, many deep neural network (DNN) methods, which obviate the need for feature engineering, have been proposed [5, 7–9, 13]. DNN methods automatically extract effective features for score prediction using a dataset of graded short answers, and have achieved state-of-the-art scoring accuracy[5, 7–9, 13]. However, further improvement of the accuracy of these methods is required, especially for high-stakes and large-scale examinations because even a slight scoring error will have a large effect on many test-takers.

To improve scoring accuracy, we propose a new ASAG method that combines a conventional DNN model and an item response theory (IRT) model [6]. We

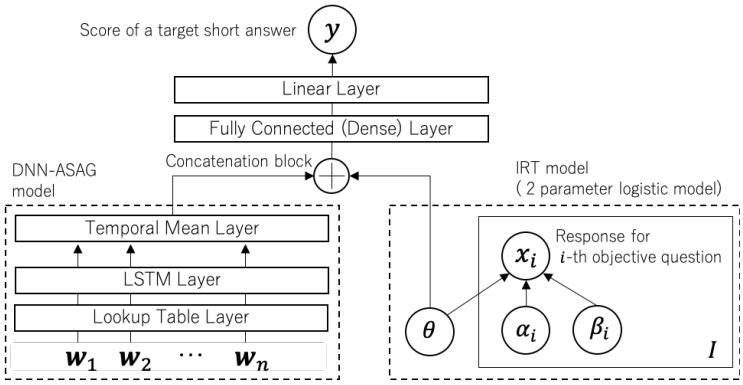

Fig. 1. Architecture of the proposed method.

focus on short-answer questions given as a part of a test including objective questions. Because a test measures a particular ability, we can assume that short-answer questions and objective questions on the same test measure similar abilities. Thus, estimating the test-takers' ability from the objective questions should be useful for short-answer grading. Based on this assumption, our method incorporates the test-taker's ability, which is estimated using an IRT model from his/her true-false responses for objective questions, into a DNN-ASAG model. Our method is formulated as a DNN framework that predicts a target short-answer score by jointly using the IRT-based ability estimate and a distributed representation of the short-answer text as obtained from an intermediate layer of a DNN-ASAG model. Although the proposed method is suitable for any DNN-ASAG model, we implement it with the most standard long short-term memory (LSTM) ASAG model [9]. The effectiveness of our model is evaluated by using data from an actual experiment. To our knowledge, this is a new approach that focuses on using responses to objective questions to grade short answers.

2 Proposed method

The architecture of the proposed method is shown in Fig. 1. The method transforms the word sequence in a given short answer to a fixed-length hidden vector $\boldsymbol{M}$ through *a lookup table layer*, *a LSTM layer*, and *a temporal mean layer*, as in the conventional LSTM ASAG model [9]. Here, *the lookup table layer* transforms each word in a given short answer to a word embedding representation, *the LSTM layer* transforms the embedded word sequence to a sequence of hidden vectors that capture the long-distance dependencies of the words at each time step, and *the temporal mean layer* averages the outputs of the LSTM layer to produce a fixed-length hidden vector $\boldsymbol{M}$, which can be regarded as a distributed representation of a given short-answer text.

The concatenation block, a newly added component in this method, concatenates the distributed text representation $\boldsymbol{M}$ and an IRT-based test-taker's ability

θ which is estimated from his/her true-false responses to objective questions offered together with the short-answer question during the same examination. We use the two-parameter logistic IRT model that defines the probability of a test-taker answering correctly for objective question i as $(1 + \exp[-\alpha_i(\theta - \beta_i)])^{-1}$, where θ is the test-taker's ability, and α_i and β_i are discrimination and difficulty parameters of question i.

The fully connected (dense) layer projects the concatenated vector $\boldsymbol{M'} = [\boldsymbol{M}, \theta]$ to a lower-dimensional hidden vector using a fully connected feedforward neural network. This layer is also newly added in this study to capture the non-linear relation between the test-takers' abilities and short-answer scores.

Finally, *the linear layer* projects the output of the fully connected layer to a scalar value in the range [0, 1] by using the sigmoid function $\sigma(\boldsymbol{W}\boldsymbol{M'} + b)$, where $\boldsymbol{W}$ is the weight matrix and b is the bias.

The model training is conducted by back-propagation with the mean squared error loss function using the training dataset, in which the scores are normalized to the [0, 1] scale. During the prediction phase, the predicted scores are rescaled to the original score range. For the IRT parameter estimation, we use a Markov chain Monte Carlo algorithm [14, 15].

3 Experiments

This section demonstrates the effectiveness of the proposed method by using real data. For this experiment, we used response data from a Japanese reading comprehension test developed by Benesse Educational Research and Development Institute, Japan. This dataset comprises responses given by 511 test-takers (Japanese university students) to three short-answer questions and true-false responses for 44 objective questions. Scores for the short answers were provided by expert raters using three rating categories for two evaluation viewpoints. The total score of the two evaluation viewpoints was also given.

Using the data, we conducted five-fold cross validation to evaluate the Pearson's correlation between the true scores and predicted scores for each evaluation viewpoint and the total score. For model training, the dimensions of the word embedding, the LSTM layer, and the fully connected layer were set to 50, 300, and 50, respectively. The mini-batch size and maximum epochs were 32 and 50, respectively. The dropout probabilities for the lookup table layer and the temporal mean layer were 0.5. The recurrent dropout probability for the LSTM layer was set to 0.1. This experiment was conducted for the proposed method and the conventional method. Furthermore, to evaluate effectiveness of the fully connected (dense) layer, we also conducted the experiment for the proposed method without the dense layer and the conventional method with the dense layer.

Table 1 shows the results. The *Score1* and *Score2* columns indicate the results for the two evaluation viewpoints in each question; the *Total* column indicates the results for the sum of the two viewpoints' scores; and the *Avg.* column shows the averaged performance for each method. * indicates that the averaged

Table 1. Experimental results

	Question 1			Question 2			Question 3			
	Score 1	Score 2	Total	Score 1	Score 2	Total	Score 1	Score 2	Total	Avg.
Conventional	0.561	0.875	0.604	0.910	0.868	0.815	0.719	0.737	0.694	0.754
with dense	0.568	0.882	0.612	0.909	0.874	0.823	0.715	0.758	0.713	0.762
Proposed	0.576	0.887	0.621	0.912	0.876	0.828	0.710	0.743	0.708	0.762 *
w/o dense	0.573	0.873	0.597	0.911	0.865	0.810	0.719	0.733	0.673	0.751

performance of the method is higher than that of the conventional method at the 1% significance level by the paired t-test.

The table shows that the proposed method has better performance than the conventional method in almost all cases, and the averaged performance of the proposed method is also significantly higher. These results suggest that the proposed method is effective in improving the scoring accuracy. The table also shows that the performance tends to decrease when the dense layer is omitted from the proposed method. Moreover, when the dense layer is added to the conventional method, the performance tends to increase. These results suggest that the incorporation of the fully connected dense layer improves the accuracy. Comparing the proposed method and the conventional method with the dense layer shows that the proposed method provides higher performance in all cases except for *Question 3*, validating the effectiveness of incorporating the IRT-based ability. The drop in performance for *Question 3* might be caused by disagreement between the distribution of IRT ability and that of the observed score. We confirmed that *Question 3* has a strongly skewed score distribution in which the highest score category is overused, whereas the IRT ability follows a normal distribution [4]. Note that test items with strongly skewed score distributions are generally inappropriate because they do not distinguish the ability of test-takers well. Thus, we conclude that incorporating ability values improves the scoring accuracy when target short-answer questions measure ability well.

4 Conclusion

This study proposed a new DNN-ASAG method that integrates the ability of test-takers estimated from true-false responses for objective questions using IRT. An experiment using real data suggested that incorporating ability improves scoring accuracy when a target short-answer question can measure ability well. In future work, we plan to examine the behavior of the proposed method in more detail by applying it to various datasets. We will also examine the potential for scoring bias that might arise from the use of true-false responses.

Acknowledgment

This work was supported by JSPS KAKENHI 17H04726 and 17K20024. We thank Yuki Doka and Yoshihiro Kato at Benesse Educational Research and Development Institute for permission to use the actual data.

References

1. Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education **25**(1), 60–117 (2015)
2. 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: Proceedings of the International Conference on Artificial Intelligence in Education. pp. 70–85 (2018)
3. Heilman, M., Madnani, N.: ETS: Domain adaptation and stacking for short answer scoring. In: Proceedings of the International Workshop on Semantic Evaluation. pp. 275–279 (2013)
4. van der Linden, W.J.: Handbook of Item Response Theory, Volume One: Models. CRC Press (2016)
5. Liu, T., Ding, W., Wang, Z., Tang, J., Huang, G.Y., Liu, Z.: Automatic short answer grading via multiway attention networks. In: Proceedings of the International Conference on Artificial Intelligence in Education. pp. 169–173 (2019)
6. Lord, F.: Applications of item response theory to practical testing problems. Erlbaum Associates (1980)
7. Lun, J., Zhu, J., Tang, Y., Yang, M.: Multiple data augmentation strategies for improving performance on automatic short answer scoring. In: Proceedings of the Association for the Advancement of Artificial Intelligence (2020)
8. Mizumoto, T., Ouchi, H., Isobe, Y., Reisert, P., Nagata, R., Sekine, S., Inui, K.: Analytic score prediction and justification identification in automated short answer scoring. In: Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics. pp. 316–325 (2019)
9. Riordan, B., Horbach, A., Cahill, A., Zesch, T., Lee, C.M.: Investigating neural architectures for short answer scoring. In: Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics. pp. 159–168 (2017)
10. Saha, S., Dhamecha, T.I., Marvaniya, S., Sindhgatta, R., Sengupta, B.: Sentence level or token level features for automatic short answer grading?: Use both. In: Proceedings of the International Conference on Artificial Intelligence in Education. pp. 503–517 (2018)
11. Sakaguchi, K., Heilman, M., Madnani, N.: Effective feature integration for automated short answer scoring. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1049–1054 (2015)
12. Sultan, M.A., Salazar, C., Sumner, T.: Fast and easy short answer grading with high accuracy. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1070–1075 (2016)
13. Sung, C., Dhamecha, T.I., Mukhi, N.: Improving short answer grading using transformer-based pre-training. In: Proceedings of the International Conference on Artificial Intelligence in Education. pp. 469–481 (2019)
14. Uto, M.: Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In: Proceedings of the International Conference on Artificial Intelligence in Education. pp. 494–506 (2019)
15. Uto, M., Ueno, M.: Item response theory for peer assessment. IEEE Transactions on Learning Technologies **9**(2), 157–170 (2016)