Automatic scoring of short answers using justification cues estimated by BERT

Shunya Takano, Osamu Ichikawa
Faculty of Data Science, Shiga University
Hikone, Japan
pppublic2020@gmail.com, osamu-ichikawa@biwako.shiga-u.ac.jp

Abstract

Automated scoring technology for short-answer questions has been attracting attention to improve the fairness of scoring and reduce the burden on the scorer. In general, a large amount of data is required to train an automated scoring model. The training data consists of the answer texts and the scoring data assigned to them. It may also include annotations indicating key word sequences. Many previous studies have created models with large amounts of training data specific to each question. This paper aims to achieve equivalent performance with less training data by utilizing a BERT model that has been pre-trained on a large amount of general text data not necessarily related to short answer questions. On the RIKEN dataset, the proposed method reduces the training data from the 800 data required in the past to about 400 data, and still achieves scoring accuracy comparable to that of humans. Annotating 400 data is still costly, but it is beneficial to reduce the number of data needed.

1 Introduction

Automatic short answer scoring (SAS) system using natural language processing technology has several advantages, such as the immediate return of scoring results and the ability to submit answers from any location over networks. To realize such interactive learning, a lot of research has been done on ASAP-SAS data. Assuming amount of scored short answers are available as training data, semantic similarity (Sultan et al. 2016) or machine learning (Zhao et al. 2017) is used for the score prediction. Also, as an attempt using deep learning, CNN and LSTM have been configured on top of word embedding to predict the holistic score directly (Riordan et al. 2017; Taghipour et al. 2016).

Unfortunately, predicting holistic scores directly from word sequences is not very promising because there are too big a leap between words and scores. With this background, RIKEN Center for AIP provided SAS dataset with analytic scores and annotations (justification cues) as well as holistic scores in public to help with research activities. The dataset includes sample responses from 2,100 students for each of the six readings comprehension test prompts (RIKEN 2020). RIKEN Center also developed automatic scoring technology using deep learning for the dataset. Mizumoto et al. proposed a bidirectional LSTM model integrating a supervised attention mechanism estimating the justification cue for scoring (Mizumoto et al. 2019). The model was evaluated with various sizes of training data. It is reported that approximately 800 training data per question are needed to achieve the same accuracy as humans. However, we know it is difficult to prepare 800 training data manually in actual schools.

Therefore, we consider using BERT model (Devlin et al. 2019) pre-trained with a large amount of general text not necessarily related to short answer questions, so to reduce the amount of specific training data required. Several research institutes provide pre-trained BERT models. They are well-trained with huge general corpus and supposed to be fine-tuned with small amount of specific corpus.

Instead of using supervised attention in Mizumoto et al. 2019, this study uses BERT to annotate word sequences as the justification cues. The justification identification model is created by fine-tuning one of the pre-trained BERT models with a specific data set.

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 8 - 13

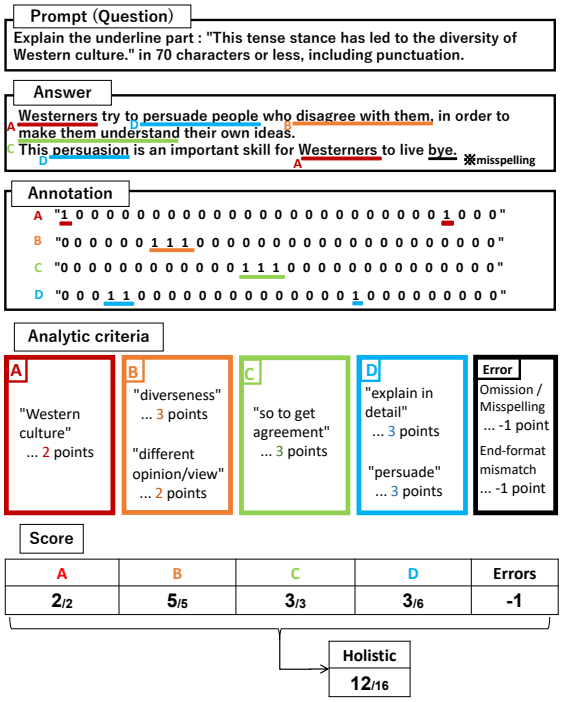

Figure 1: Example of data in Riken dataset (Translated from Japanese). Note this study does not consider point deductions based on the "error" criterion.

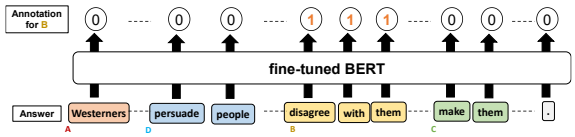

Figure. 2: Image of the justification identification model.

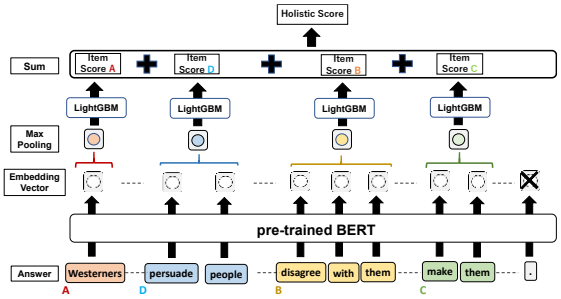

Figure 3: Image of the score prediction model.

2 RIKEN dataset for short answer assessment

Our study compares the performance and the required data size with the previous method by Mizumoto et al. 2019. Therefore, the same RIKEN dataset was used. This data was obtained by annotating and scoring the answers to Japanese writing questions in the mock examinations for high school students conducted by Yoyogi Seminar from 2014 to 2015 according to a predefined rubric. Answers are approximately 50 to 70 characters long (in Japanese). As shown in Figure 1, each answer is annotated with a "1" for words that contribute to the score and a "0" for words that do not. The overall score is calculated as the sum of the individual scores. In the example shown in Figure 1, there are four analytic criteria, A, B, C, and D, and annotations are assigned to each of them. Each item is then scored using this annotation as a justification cue. The overall score is calculated by summing the item scores and subtracting points for spelling errors and bad sentence endings etc.

3 Proposed Method

3.1 Justification prediction

The RIKEN dataset includes justification cues (annotations) indicating the words in the answer text that support each of the analytic criteria, such as A, B, C, and D in Figure 1. Justification cues were integrated in the scoring model as supervised attention in the previous method (Mizumoto et al. 2019). As our approach uses them to select embedding vectors, they must be explicitly predicted by a separate model.

Figure 2 shows our proposed justification prediction model. The pre-trained BERT model is fine-tuned with the answer text as input, and the "0" and "1" annotations as teacher signal. In this figure, taking scoring criterion B as an example, "1" annotations are output for words that support B.

Note that prompt phrases are not used to predict annotations. The performance of the justification identification model solely depends on the quality of the annotated answer text used for training.

3.2 Analytic score and holistic scores predictions

Figure 3 shows the proposed model. First, the answer is divided into units of strings called tokens, which are converted into IDs and passed through the pre-trained BERT model. The tokens are then converted into 768-dimensional embedding vectors. It should be noted that the embedding vector of BERT is context-aware unlike the one of Word2Vec and etc. This is one of the advantages of using BERT. For each scoring criterion, we collect the embedding vectors of the word tokens that are annotated with "1", which indicating these words

support the analytic criteria. A new 768-dimensional vector is generated by taking the maximum value for each dimension of the collected vectors. Using the vectors as features, analytic scores for each item are predicted by the respective LightGBM model trained on the same data used in the justification identification model.

If the annotations are all "0", the score for the corresponding item is set to 0 because there is no vector to feed the score prediction model. Finally, the holistic score is calculated by summing up all the item scores.

4 Experiments

The following experiments were conducted to evaluate the performance of the proposed method. Our experiments used the Japanese pre-trained BERT model published by the Inui-Suzuki Laboratory at Tohoku University (Inui Laboratory 2021).

4.1 Settings

RIKEN Dataset for short answer assessment was used for the experiments. As in the previous study (Mizumoto et al. 2019), we used 6 out of 9 test prompts. They are denoted by Q1 through Q6 in the tables in this paper. There are 2100 answer sheets for each prompt. The holistic score was calculated by summing up all the item scores. In this study, deduction for errors of misspellings, omissions, sentence endings etc. was not considered.

To evaluate the performance of the models, we created several test cases with different sizes of training data such as 100, 200, 400, 800 and 1600. For example, in 100-training case, 100 answers were used as training data and the remaining 2000 answers as test data. Similarly, in 400-training case, 400 answers were used as training data and the remaining 1700 answers as test data. Each case consisted of five sets of training data selected to have as little overlap as possible between the sets, and performance was measured by the average of the five sets.

[1] Quadratic Weighted Kappa (QWK) is an evaluation metric for multi-class classification. It takes a value from 0 to 1, with a higher value indicating a better fit of the prediction. In this study, we convert the predicted overall scores into integers by rounding off fractions and treat the integer scores as classes for QWK.

	Q1	Q2	Q3	Q4	Q5	Q6
100 train	0.97	0.96	0.91	0.89	0.93	0.94
200 train	0.98	0.97	0.94	0.91	0.95	0.96
400 train	0.98	0.98	0.95	0.93	0.96	0.96
800 train	0.99	0.98	0.96	0.94	0.96	0.97
1600 train	0.99	0.98	0.95	0.95	0.97	0.97
Human	0.96	0.94	0.76	0.84	0.82	0.90

Table 1: QWK with correct justification cue given.

	Q1	Q2	Q3	Q4	Q5	Q6
100 train	0.77	0.59	0.31	0.61	0.65	0.63
200 train	0.81	0.71	0.38	0.66	0.70	0.70
400 train	0.85	0.77	0.44	0.71	0.74	0.74
800 train	0.87	0.82	0.47	0.73	0.77	0.77
1600 train	0.90	0.84	0.53	0.75	0.79	0.79

Table 2: QWK without using justification cue.

	Precision	Recall	F-measure
Current	0.848	0.895	0.866
Mizumoto	0.837	0.703	0.758

Table 3: Performance of justification identification (100 training data case)

100 train		Q1	Q2	Q3	Q4	Q5	Q6
Analytic criteria	A	0.970	0.928	0.867	0.936	0.839	0.809
	B	0.912	0.914	0.840	0.859	0.883	0.891
	C	0.937	0.973	0.746	0.885	0.954	0.819
	D	0.922	0.844	0.468	—	—	—

Table 4: F-measure of annotation (100 train).

400 train		Q1	Q2	Q3	Q4	Q5	Q6
Analytic criteria	A	0.982	0.954	0.902	0.965	0.869	0.910
	B	0.940	0.949	0.890	0.872	0.896	0.923
	C	0.956	0.980	0.820	0.910	0.960	0.853
	D	0.942	0.869	0.724	—	—	—

Table 5: F-measure of annotation (400 train).

4.2 (Preliminary Experiment) Automatic scoring with / without correct justification cues

To investigate the upper bound of the performance of the score prediction model shown in Section 3.2, we predict the item score and holistic scores with using the correct justification cues given by the dataset. Also, we investigated the lower bound without using any justification cues. Quadratic Weighted Kappa (QWK)[1] (Cohen 1960) was used as evaluation metrics for the holistic score, and the mean values calculated on the five sets are shown in Table 1 and Table 2. Table 1 also shows the human scoring accuracy in QWK, which was reported in Mizumoto et al. 2019.

4.3 (Experiment 1) Justification identification

We evaluated the performance of the justification identification model shown in Section 3.1. Although Figure 2 shows only scoring criterion B, we trained 21 BERT models to predict the annotations for all analytic criteria for each test prompt.

The BERT models were prepared by fine-tuning the pre-trained BERT models with the number of epochs set to 10, batch size set to 16, optimization algorithm set to Adam, and loss function set to cross-entropy function.

Table 3 outlines the performance of justification identification for the case of 100-training data. It also shows the supervised attention case reported in Mizumoto et al. 2019 which also reports 100-training data case. Table 4 and Table 5 provide breakdowns of all analytic criteria in the 100-training case and the 400-training case. Please note each of the six test prompts, from Q1 to Q6, has its own analytic criteria from A to D (or C).

4.4 (Experiment 2) Automatic scoring using automatically predicted justification cues

We evaluated the performance by combining both models shown in Section 3. The justification cues were predicted by the model shown in Section 3.1, and the item and holistic scores were predicted by the model shown in Section 3.2, using the predicted justification cues and embedding vectors. QWK was used as evaluation metrics, and the mean values[2] of the five sets of the metric are shown in Table 6.

4.5 Discussion of the experimental results

As shown in Table 1, given the correct justification cues, the accuracy in QWK of automatic scoring by the proposed model is much higher than human scoring for all questions, even when using only 100 training data. On the other hand, accuracy was poor when justification cues were not used. This indicates that justification cues are critically important in SAS, especially in our model proposed in Section 3.2.

[2] The values have been updated since our last report in domestic meeting of FIT 2021, due to the calculation errors Also, we found increasing epochs from 3 to 10 in fine-tuning of BERT significantly improved the accuracy of justification cue prediction.

	Q1	Q2	Q3	Q4	Q5	Q6
100 train	0.94	0.88	0.65	0.80	0.82	0.83
200 train	0.96	0.91	0.74	0.83	0.85	0.87
400 train	0.97	0.92	0.77	0.85	0.86	0.89
800 train	0.97	0.95	0.80	0.87	0.88	0.91
1600 train	0.98	0.95	0.83	0.88	0.88	0.92
Human	0.96	0.94	0.76	0.84	0.82	0.90

Table 6: QWK with predicted justification cue.

	100	200	400	800	1600
No cues	0.590	0.659	0.706	0.737	0.766
Given cues	0.934	0.950	0.959	0.965	0.969
Predicted cues	0.820	0.857	0.877	0.894	0.906
Mizumoto	0.776	0.827	0.856	0.876	0.892

Table 7: QWK summaries of all experiments and references for training data of various sizes.

With respect to the accuracy of justification identification, Table 3 shows that our fine-tuned BERT model can identify cues much better than the supervised attention model reported in Mizumoto et al. 2019. Table 5 provides the details in F-measure in 400-training data case. The BERT model worked well, with high accuracy on most items. One exception is criterion D of Q3, which concerns human emotions such as "frustration" and "distress", unlike the other analytic criteria. Even BERT may not be able to properly translate human emotions into numeric vectors.

The performance of our proposed method integrating the two models is shown in Table 6. With 400 training data, the QWK values are quite close to human scoring. This means our justification identification model successfully selected the BERT embedding vectors that form the input to the analytic scoring model of LightGBM. However, comparing Table 6 and Table 1, the upper bound results using given correct justification cues are still much better. This suggests further refinement in justification identification model would be desirable in the future.

Table 7 summarizes the experimental results for various sizes of training data. Given the correct justification cues, the performance degradation when training data is small is very small. As the proposed method improved cue prediction, it performed better than the comparative method (Mizumoto et al. 2019), especially when training data was small, such as 100 or 200 training data.

5 Conclusion

This paper proposed the combined model of justification prediction and analytic scoring model. It includes fine-tuning of pre-trained BERT model that predicts justification cues (annotations), which are crucial for automatic scoring. BERT embedding vectors of annotated words are subsequently passed to LightGBM model (Ke et al. 2017) for scoring. The proposed model uses a BERT model that has been pre-trained with a large corpus of text in a general domain. As shown in Table 7, this helped automated scoring on specific data sets and showed that the accuracy of scoring on the RIKEN dataset can be comparable (0.88) to that of human scorers (average 0.873) with training data of only 400 answers per prompt. Compared to the comparative method (Mizumoto et al. 2019) which showed an accuracy of 0.87 with 800 answers, almost 50% reduction of training data has been achieved.

Acknowledgments

In this paper, we used "RIKEN Dataset for Short Answer Assessment" provided by RIKEN via IDR Dataset Service of National Institute of Informatics. This work was supported by JSPS KAKENHI (Grant Numbers JP19K02999).

References

ASAP-SAS. 2012. Scoring short answer essays. ASAP short answer scoring competition system description.

Md Arafat Sultan, Cristobal Salazar and Tamara Sumner. 2016. Fast and easy short answer gradingwith high accuracy. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1070–1075.

Siyuan Zhao, Yaqiong Zhang, Xiaolu Xiong, Anthony Botelho and Neil Heffernan. 2017. A memory-augmented neural model for automated grading. *Proceedings of Fourth ACM Conference on Learning @ Scale*, pages 189–192.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch and Chungmin Lee. 2017. Investigating neural architectures for short answer scoring. *Proceedings of the 12th Workshop on Building Educational Applications Using NLP (BEA)*, pages 159-168.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882-1891.

RIKEN. 2020. RIKEN Dataset for Short Answer Assessment. *Informatics Research Data Repository, National Institute of informatics.* Dataset: https://doi.org/10.32130/rdata.3.1

Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine and Kentaro Inui. 2019. Analytic score prediction and justification identification in automated short answer scoring. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 14)*, pages 316-325.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. BERT: pre training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171-4186.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye and Tie-Yan Liu. 2017. LightGBM: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, pages 3149-3157.

Inui Laboratory, Tohoku University. 2021. Pretrained Japanese BERT models. https://github.com/cl-tohoku/bert-japanese (referred on June 10, 2021.)

Jacob. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1): 37–46.