What can Short Answer Scoring Models Learn from Cross-prompt Training Data?

Hiroaki Funayama[1,2] Yuya Asazuma[1,2]
Yuichiroh Matsubayashi[1,2] Tomoya Mizumoto[2] Kentaro Inui[1,2]
[1]Tohoku University [2]RIKEN
{h.funa, asazuma.yuya.r7}@dc.tohoku.ac.jp
{y.m, kentaro.inui}@tohoku.ac.jp tomoya.mizumoto@a.riken.jp

Abstract

For the task of Automatic Short Answer Scoring（ASAS), both rubrics and reference answers differ for every single prompt which requires a need to annotate answers for each in order to construct a highly-effective scoring model. Such need to annotate answers for every prompt is costly, especially in the context of school education and online courses where only a few answers to prompts exist. In this work, we attempt to reduce this burden by first training a model for predicting scores given rubrics and answers of already annotated prompts (adaptive-pretraining). We then fine-tune the model on a small amount of data for each new prompt to be graded. Our experimental results show that adaptive-pretraining with rubrics significantly improve scoring accuracy, especially when the training data is scarce.

1 Introduction

Automatic Short Answer Scoring（ASAS) is the task of automatically scoring a given input (e.g., essays) to a prompt based on existing rubrics and reference answers [1, 2, 3, 4]. ASAS has been extensively studied as a means to reduce the burden of manually scoring student answers in both school education and large-scale examinations. Recently, the practical application of ASAS systems has gained much attention, both in school education and e-learning [5, 6, 7]. However, the annotation cost is limited in school education and online courses, making it challenging to obtain sufficient training data for developing ASAS models [8]. The data to train ASAS models must be prepared for each prompt independently, as the rubrics and reference answers are different for each prompt [9]. Those facts are a considerable barrier to the practical application of ASAS systems in these situations. However, whether cross-prompt training of ASAS models can reduce annotation costs is one of the biggest open issues in this field, and few studies have addressed it [10].

Towards answering this question, given that annotated data from other prompts cannot be used directly for training an ASAS model for a specific prompt since the rubrics and reference answers for each prompt are entirely different, we focus on the relationship between the rubrics and the answers. ASAS is a task that assigns a score when an answer implicates an expression described in the rubric. Thus, we can view ASAS as determining implication relations by flexibly matching the semantics between the rubrics and answers. We leverage already annotated prompts to make the model recognize such implication relationships flexibly. Inspired by this idea, we attempt to make a model (i.e., BERT) learn the relationship between rubrics and answers using already annotated prompts (see fig.1). We can't use them directly to train the model because the rubrics contain various information for grading answers. Therefore, we utilize key phrases [11, 12], representative examples of expressions that an answer should include to gain scores.

We train BERT on already graded prompts as predicting scores by inputting key phrase/answer pairs (name it adaptive-pretraining). Thorough the adaptive-pretraining, we expect the model will learn the relationship between the key phrase and the expressions in the answers. We then finetune BERT on the prompts to be graded. With adaptive-pretraining, we expect to build ASAS models that are more robust against expression variation, especially when the training data is scarce.

In our experiment, we examine the effectiveness of

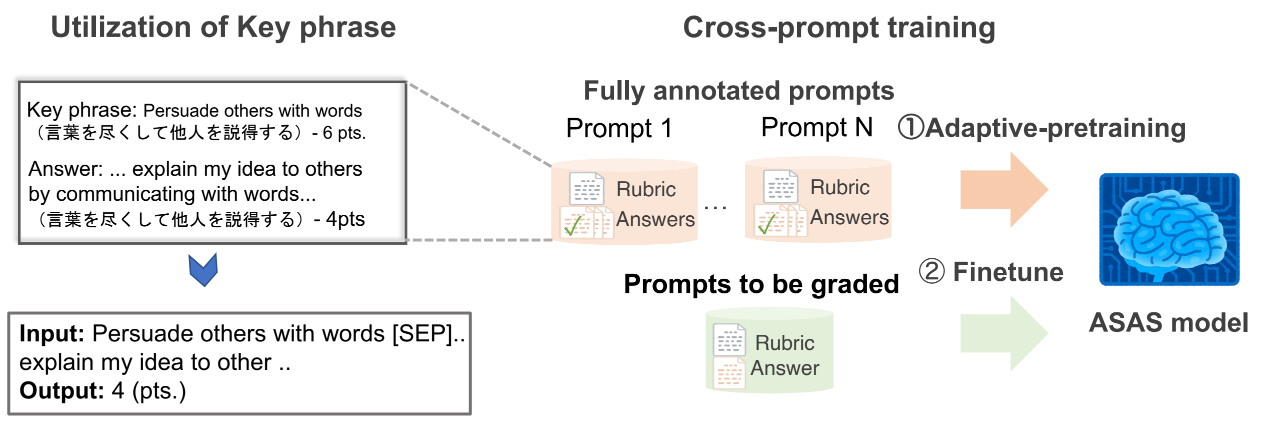

Utilization of Key phrase
Cross-prompt training
Key phrase: Persuade others with words
（言葉を尽くして他人を説得する）- 6 pts.
Answer: ... explain my idea to others
by communicating with words...
（言葉を尽くして他人を説得する）- 4pts
Input: Persuade others with words [SEP]..
explain my idea to other ..
Output: 4 (pts.)
Fully annotated prompts
Prompt 1
Prompt N
Rubric
Answers
...
Rubric
Answers
①Adaptive-pretraining
Prompts to be graded
Rubric
Answer
② Finetune
ASAS model

Figure 1 Overview of our proposed method. We input a key phrase, reference expressions, with an answer. We first adaptive-pretrain the ASAS model on already annotated prompts and then finetune the model on a prompt to be graded.

adaptive-pretrain with key phrases. Our experimental results show that the model performance improves by 0.17 at maximum in Quadratic Weighted Kappa (QWK) by adaptive-pretraining with key phrases.

In addition to our experiments, towards effective ASAS modeling which requires a significant amount of prompts and answers, we contribute 1,0000 new data annotations (20 prompts with 500 answers each) to the RIKEN dataset [11], the only Japanese dataset available for ASAS. We make our data annotations publicly available for academic purposes.

2 Method

As mentioned a priori, in this study, we attempt to train ASAS models in a cross-prompt manner in order to reduce the data required to train a model for a new prompt to be graded automatically by leveraging data from already annotated prompts. Specifically, we first train a BERT model to predict scores from pairs of answers and key phrases of the already annotated prompts (adaptive pretrain). Next, we further fine-tune BERT with the small amount of data from the prompt to be graded automatically.

2.1 Task definition

In this study, we assume that fully annotated prompts are available. We consider utilizing those annotated prompts to train models for the newly obtained prompts to be graded automatically. Let P_{known} denote the already annotated prompts and P_{target} denote the newly obtained prompts to be graded. Suppose X_p represents a set of all possible student's answers of a given prompt $p \in P$, and $\mathbf{x} \in X_p$ is an answer. The prompt has an integer score range from 0 to N, which is basically defined in rubrics. Namely, the score for each answer is selected from one of the integer set $S = \{0, ..., N\}$. Therefore, we can define the ASAS task as assigning one of the scores $s \in S$ for each given input $\mathbf{x} \in X_p$. Moreover, to construct an ASAS model means to construct a regression function m from every input of student answer $\mathbf{x} \in X$ to a score $s \in S$, that is, $m : X \rightarrow S$.

2.2 Scoring model

A typical, recent approach to constructing a mapping function m is the use of newly developed deep neural networks (DNNs). As discussed a priori, the set of scores S consists of several consecutive integers $\{0, \ldots, N\}$. Suppose D is training data that consist of a set of actually obtained student's answers $\mathbf{x}$ and its corresponding human annotated score s pairs, that is, $D = ((\mathbf{x}_i, s_i))_{i=1}^{I}$, where I is the number of training data. To train the model m, we try to minimize the Mean Squared Error (MSE) loss on training data $L_m(D)$ calculated using model m. Therefore, we can write the training process of the SAS model as the following minimization problem:

$$m = \underset{m'}{\mathrm{argmin}} \left\{ L_{m'}(D) \right\}, \quad L_m(D) = \frac{1}{|D|} \sum_{(\mathbf{x},s) \in D} (s - m(\mathbf{x}))^2, \tag{1}$$

where $m(\mathbf{x})$ represents the calculated prediction of model m given input $\mathbf{x}$. Once m is obtained, we can predict the score $\widehat{s}$ of any input (student answer) by using trained model m as $\widehat{s} = m(\mathbf{x})$.

2.3 Adaptive-pretrain with rubrics

2.3.1 Key phrase

A key phrase is a representative example of the expressions that an answer must contain in order to gain scores. In general, rubrics are difficult to utilize directly because they detail the information and properties that an answer must contain in order to score. In general, it is difficult to use rubrics for training models directly because they detail the information and expressions that an answer must contain to gain scores. Therefore, we utilize key phrases such as those shown in Figure 2.

2.3.2 Adaptive-pretrain

As described in Section 2.1, we utilize data from already annotated prompts P_{known} to train models for prompts P_{target} to be graded. For each prompt $p \in P$, there exists a key phrase k_p.We separate the key phrase k_p of prompt p and the i-th answer $x_{p,i}$ of prompt p by [SEP] as the sequence $t_{p,i} = \{k_p, [SEP], \mathbf{x_{p,i}}\}$. Then, we construct data for adaptive-pretraining as:

$$D_{adap} = \{(t_{p,i}, s_{p,i}) | p \in= P_{known}\}_{i=1}^{I} \tag{2}$$

We train the BERT-based regression model on this dataset $\mathcal{D}_{adap}$ to obtain model m_{adap}:

$$m_{adap} = \underset{m'}{\operatorname{argmin}} \left\{ L_{m'}(D_{adap}) \right\} \tag{3}$$

We refer to models trained on existing graded prompts as *adaptive-pretraining*.

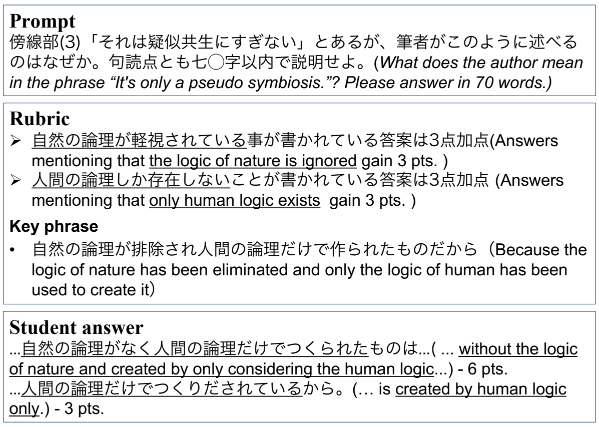
Next, we further fine-tune the adaptive-pretrained model on $p \in P_{target}$ to obtain a model m_p for prompt p.

$$m_p = \underset{m'}{\operatorname{argmin}} \left\{ L_{m'}(D_p \right\}, \tag{4}$$

3 Experiment

3.1 Dataset

We use the RIKEN dataset, the only publicly available Japanese SAS dataset[1] provided in [11]. As mentioned in Section.1, we extend the dataset to conduct this research. Each prompt in the RIKEN dataset has several scoring rubrics (i.e., analytic criterion [11]), and an answer is manually graded based on each analytic criterion independently (i.e., analytic score). Thus, following [13], we treat this analytic criterion as an individual scoring task. For simplicity, we refer to each analytic criterion as a single prompt in this paper. In this way, we consider that there are a total of 109 prompts in this dataset.



Prompt
傍線部(3)「それは疑似共生にすぎない」とあるが、筆者がこのように述べるのはなぜか。句読点とも七〇字以内で説明せよ。(*What does the author mean in the phrase "It's only a pseudo symbiosis."? Please answer in 70 words.)*

Rubric
- 自然の論理が軽視されている事が書かれている答案は3点加点(Answers mentioning that the logic of nature is ignored gain 3 pts.)
- 人間の論理しか存在しないことが書かれている答案は3点加点 (Answers mentioning that only human logic exists gain 3 pts.)

Key phrase
- 自然の論理が排除され人間の論理だけで作られたものだから（Because the logic of nature has been eliminated and only the logic of human has been used to create it）

Student answer
...自然の論理がなく人間の論理だけでつくられたものは...(... without the logic of nature and created by only considering the human logic...) - 6 pts.
...人間の論理だけでつくりだされているから。(… is created by human logic only.) - 3 pts.

Figure 2 Example of a prompt, scoring rubric, and student's answers excerpted from RIKEN dataset [11] and translated from Japanese to English. For space reasons, some parts of the rubrics and answers are omitted.

In our experiment, we used 21 prompts as P_{target} to evaluate the effectiveness of adaptive-pretraining (see Appendix for detailed information regarding 3). For adaptive-pretraining, we used all remaining 88 prompts consisting of 480 answers per prompt for training the model and 20 answers per prompt as devset.

3.2 Setting

As described in Section 2.2, we used pretrained BERT [14] as the encoder for the automatic scoring model and use the vectors of CLS tokens as feature vectors for predicting answers [2]

Similar to previous studies [11, 15, 2], we use a Quadratic Weighted Kappa (QWK) [16], a de facto standard evaluation metric in ASAS, in the evaluation of our models. The scores were normalized to a range from 0 to 1 according to previous studies [15, 11]. QWK was measured by re-scaling when evaluated on the test set. We train a model for 5 epochs in the adaptive-pretraining process. We then fine-tune the adaptive-pretrained model for 10 epochs. In the setting without adaptive-pretraining process, we fine-tune the model for 30 epochs. We computed the QWK of the dev set at the end of each epoch

[1] https://aip-nlu.gitlab.io/resources/sas-japanese

[2] We used pretrained BERT models from https://huggingface.co/bert-base-uncased for English and https://github.com/cl-tohoku/bert-japanese for Japanese.

Table 1 QWK and standard deviation of four settings; with and without adaptive-pretraining, and with and without rubrics (keyphrase). In the adaptive-pretraining phase, we use 88 prompts, 480 answers per prompt. We change the amount of data for finetuning as 25, 50, 100, and 200.

# data for finetune	w/ Adaptive-pretraining		w/o Adaptive-pretraining	
	w/ rubric	w/o rubric	w/ rubric	w/o rubric
25	0.67 ± 0.02	0.47 ± 0.03	0.51 ± 0.03	0.50 ± 0.01
50	0.74 ± 0.01	0.62 ± 0.02	0.64 ± 0.02	0.64 ± 0.01
100	0.78 ± 0.01	0.70 ± 0.02	0.73 ± 0.02	0.73 ± 0.01
200	0.81 ± 0.01	0.77 ± 0.01	0.80 ± 0.01	0.79 ± 0.01

in fine-tuning process and evaluated the test set using the parameters with the maximum QWK.

3.3 Results

Table 2 QWK and standard deviation when the total number of answers used for adaptive-pretrain is fixed at 1,600 and the number of questions used is varied from 5, 10, 20, 40, 80. 50 training data were used for finetuning.

#prompt	#data per prompt	QWK
5	320	0.68 ± 0.02
10	160	0.68 ± 0.02
20	80	0.74 ± 0.01
40	40	0.74 ± 0.01
80	20	0.74 ± 0.00

We first compared the performance with and without adaptive pretraining and with and without scoring criteria. Here, similar to [11], we experimented with 25, 50, 100, and 200 training data instances in the fine-tuning phase. The results are shown in Table 1. First, we can see that adaptive pretraining without key phrases does not improve the model performance. Similarly, using only key phrases without adaptive pretraining does not improve scoring accuracy. QWK improves significantly only when key phrases are used and when adaptive-pretrain is performed. The gain was notably large when the training data was scarce, with a maximum improvement of about 0.17 in QWK when using 25 answers for fine-tuning. On the other hand, the performance did not improve when we used 200 answers in training, which indicates that adaptive-pretraining does not benefit when sufficient training data is available. Furthermore, it is also suggested that the adaptive pretraining with key phrases can reduce the required training data by half while maintaining the same performance. Note that the results without adaptive-pretrain are comparable to the results of the baseline model shown in [11].

Impact of the number of prompts used for adaptive-pretraining Next, we examined how changes in the number of prompts affect adaptive-pretrain: we fixed the total number of answers used for the adaptive-pretrain at 1,600 and varied the number of prompts between 5, 10, 20, 40, and 80. We performed fine-tuning using 50 answers for each prompt. The results are shown in Table 2. The QWK is 0.68 when the number of prompts is 5 or 10, indicating that the effectiveness of adaptive-pretraining is inferior when the number of prompts is less than 20. This suggests that a sufficient number of prompts are required for effective adaptive-pretraining. It also suggests that increasing the number of prompts is more effective for adaptive-pretraining than increasing the number of answers per prompts.

4 Conclusion

The limited cost of annotation for data has been a major obstacle in deploying ASAS systems into school education and online learning courses. To tackle this problem, we considered utilizing already annotated prompts. Specifically, we first performed adaptive-pretraining for a BERT-based regression model using the answers and key phrases of the annotated questions. We then further fine-tuned the BERT model with a small amount of data on the prompt we want to grade automatically.

Experimental results showed that adaptive-pretraining with key phrases greatly improves the performance of the model, especially when the training data is scarce. We also discovered that adaptive-pretraining can reduce the amount of required training data by half while maintaining the same performance.

Acknowledgement

We are grateful to Dr. Paul Reisert for their writing and editing assistance. This work was supported by JSPS KAKENHI Grant Number 22H00524, JP19K12112, JST SPRING, Grant Number JPMJSP2114. We also appreciate Takamiya Gakuen Yoyogi Seminar for providing the valuable data.

References

[1] Claudia Leacock and Martin Chodorow. C-rater: Automated Scoring of Short-Answer Questions. **Computers and the Humanities**, Vol. 37, No. 4, pp. 389–405, 2003.

[2] Michael Mohler, Razvan Bunescu, and Rada Mihalcea. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 752–762, 2011.

[3] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. Fast and easy short answer grading with high accuracy. In **NAACL-HLT**, pp. 1070–1075, San Diego, California, June 2016. Association for Computational Linguistics.

[4] Surya Krishnamurthy, Ekansh Gayakwad, and Nallakaruppan Kailasanathan. Deep learning for short answer scoring. **International Journal of Recent Technology and Engineering**, Vol. 7, pp. 1712–1715, 03 2019.

[5] Yaman Kumar, Swati Aggarwal, Debanjan Mahata, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. Get it scored using autosas — an automated system for scoring short answers. In **AAAI/IAAI/EAAI**. AAAI Press, 2019.

[6] Shourya Roy, Sandipan Dandapat, Ajay Nagesh, and Y. Narahari. Wisdom of students: A consistent automatic short answer grading technique. In **Proceedings of the 13th International Conference on Natural Language Processing**, pp. 178–187, Varanasi, India, December 2016. NLP Association of India.

[7] Xiaoming Zhai. Practices and theories: How can machine learning assist in innovative assessment practices in science education. **Journal of Science Education and Technology**, Vol. 30, No. 2, pp. 139–149, Apr 2021.

[8] Torsten Zesch, Michael Heilman, and Aoife Cahill. Reducing annotation efforts in supervised short answer scoring. In **Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 124–132, Denver, Colorado, June 2015. Association for Computational Linguistics.

[9] Steven Burrows, Iryna Gurevych, and Benno Stein. The eras and trends of automatic short answer grading. **International Journal of Artificial Intelligence in Education**, Vol. 25, No. 1, pp. 60–117, 2015.

[10] Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. Survey on automated short answer grading with deep learning: from word embeddings to transformers, 2022.

[11] Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine, and Kentaro Inui. Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring. In **BEA**, pp. 316–325, 2019.

[12] Tianqi Wang, Hiroaki Funayama, Hiroki Ouchi, and Kentaro Inui. Data augmentation by rubrics for short answer grading. **Journal of Natural Language Processing**, Vol. 28, No. 1, pp. 183–205, 2021.

[13] Hiroaki Funayama, Tasuku Sato, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. Balancing cost and quality: An exploration of human-in-the-loop frameworks for automated short answer scoring. In Maria Mercedes Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova, editors, **Artificial Intelligence in Education**, pp. 465–476, Cham, 2022. Springer International Publishing.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **NAACL-HLT**, pp. 4171–4186, June 2019.

[15] Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. Investigating neural architectures for short answer scoring. In **BEA**, pp. 159–168, 2017.

[16] Jacob Cohen. Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. **Psychological bulletin**, Vol. 70, No. 4, pp. 213–220, 1968.

A List of prompts used in the evaluation

In this study, we considered the use of already annotated prompts for constructing ASAS models for new prompts to be graded. Therefore, in the experiments, we divided all prompts into two categories, prompts used for adaptive pretraining (already annotated prompts) and prompts used for the evaluation (prompts to be graded). We shows the list of prompts used for the evaluation in table 3. We used all prompts except those in this table for adaptive-pretrain.

Table 3 List of the prompts and analytic criterion used for the evaluation.

prompt	analytic criterion
Y14_1-2_1_3	A, B, C, D
Y14_1-2_2_4	A, B, C, D
Y14_2-1_2_3	A, B, C, D
Y14_2-1_1_5	A, B, C
Y14_2-2_1_4	A, B, C
Y14_2-2_2_3	A, B, C