Fully Automated Short Answer Scoring of the Trial Tests for Common Entrance Examinations for Japanese University

Haruki Oka[1(✉)], Hung Tuan Nguyen[2], Cuong Tuan Nguyen[2], Masaki Nakagawa[2], and Tsunenori Ishioka[3]

[1] Recruit Co., Ltd., Tokyo, Japan
haruki_oka@r.recruit.co.jp
[2] Tokyo University of Agriculture and Technology, Tokyo, Japan
nakagawa@cc.tuat.ac.jp
[3] The National Center for University Entrance Examinations, Tokyo, Japan
tunenori@rd.dnc.ac.jp

Abstract. Studies on automated short-answer scoring (SAS) have been conducted to apply natural language processing to education. Short-answer scoring is a task to grade the responses from linguistic information. Most answer sheets for short-answer questions are handwritten in an actual educational setting, which is a barrier to SAS. Therefore, we have developed a system that uses handwritten character recognition and natural language processing for fully automated scoring of handwritten responses to short-answer questions. This is the most extensive scoring data for responses to short-answer questions, and it may be the largest in the world. Applying the Cohen's kappa coefficient to the graded evaluations, the results show 0.86 in the worst case, and approximately 0.95 is recorded for the remaining five question answers. We observe that the fully automated scoring system proposed in our study can also score with a high degree of accuracy comparable to that of human scoring.

Keywords: Short answer scoring · Natural language processing · Handwritten character recognition

1 Introduction

Considering the current educational field, descriptive questions are often introduced to properly evaluate the abilities developed in linguistics. Moreover, to improve the scoring process's efficiency and stability, the effective use of computers and artificial intelligence has recently been increasing. There are approximately two types of descriptive questions: "essays without a correct answer" and "short-answer questions with correct answers." Many systems have been developed and have been practicalized for essays, especially in the United States.

H. Oka—Work done while at The University of Tokyo.

M. M. Rodrigo et al. (Eds.): AIED 2022, LNCS 13355, pp. 180–192, 2022.
https://doi.org/10.1007/978-3-031-11644-5_15

Some of the systems include the e-rater [2], IntelliMetric [19], intelligent essay assessors (IEA) [7], and CRASE [13]. Although the importance of short-answer questions has been recognized, various technical issues remain unsolved, such as semantic incomprehension.

On the other hand, short-answer questions are often used in several cases. Because short-answer questions are widely regarded as more orthodox, authentic, and reliable than the traditional multiple-choice tests [6], they have the potential to be used if the technical challenges for scoring are overcome. Automated short-answer scoring (SAS) techniques for English language have undergone technical improvements. Since the proposal of SAS that uses deep learning, its (SAS) performance has improved [1,5,17]. Particularly, SAS was devised using a massive transformer-based language model [8,12,21,22]. The demand of SAS is immeasurable and is not limited to new tests. Therefore, recent studies on SAS for practical purposes in Japan use data from actual mock examinations [8,15].

However, these studies have two unresolved problems. First, SAS requires additional manual work. It takes time and effort to convert handwritten data into electronic media because most of the descriptive answers in the educational domain are handwritten. The conventional SAS method aims to reduce the effort involved in scoring and requires extra effort. Furthermore, annotations were added as a guide for scoring to ensure accuracy. Considering its practical use in education, SAS requires improvements to eliminate these efforts. We have produced a fully automated scoring system that reliably eliminates data processing (such as annotations) and converts handwritten responses into text data. Second, the data handled in actual educational settings were too few to be verified on a large scale. When considering the privacy viewpoint, the amount of data was limited, and the verification was limited to a small scale. We conducted an experiment using data from a nationwide test and clarified that we could guarantee high prediction accuracy, even with large-scale data from actual educational settings.

The contributions of our research are as follows:

- We have developed a fully automated scoring system for handwritten responses, making it possible to grade many handwritten responses with high accuracy cost effectively.
- Large-scale data collected from two trial tests of entrance examinations nationwide were used to verify the practicality of the method in education.

Section 2 describes the large dataset used for the trial test of Japanese common entrance examinations. Section 3 explains the handwriting recognition technology and the scoring model used. The recognition evaluation criteria were also added. Section 4 presents the evaluation results, Sect. 5 describes the ablation studies, and Sect. 6 concludes the paper.

2 Trial Test Dataset for University Common Entrance Examinations

2.1 Overview

We used the written answers in Japanese in the trial test for the university common entrance examination conducted in 2017 and 2018. These exams are for national and private Japanese universities and are jointly conducted by the National Center for University Entrance Examinations, an independent administrative organization in Japan.

Approximately 500,000 examinees nationwide take the exams annually. These exams are considered essential for admission to national universities. Moreover, many leading private universities base their admission on these exams. Japanese exam questions comprise only first-appearing questions and are conducted once annually. While, SAT and ACT use test items repeatedly, and carry out many times a year.

We used trial test data for university common entrance exams conducted in 2017 and 2018. The test questions were prepared in a manner similar to the production, and the quality of the test questions was rigorously examined. Regarding the trial test, items on the national language (i.e., Japanese), mathematics, geography, history, civics, science, and foreign languages (only in 2018) were included. Descriptive questions were used only in Japanese and mathematics. Approximately 38% of high schools in Japan participated in this trial test; nonetheless, candidates did not have to take all the subjects. We analyzed the national language, which was taken by approximately 60,000 people. This is an unprecedented number of short-answer data for analysis.

2.2 Short-Answer Questions

The national language test in the trial test consisted of five test sets, known as the item bundles. One of these questions was a short-answer question. The test set consisted of three test questions. In 2017, these three test questions needed to be answered within 50, 25, and 120 characters, respectively. In 2018, the answers were to be of 30, 40, and 120 characters. Two Japanese characters are roughly equivalent to one English word. Figure 1 demonstrates a short-answer question administered in 2018.

3 Method

3.1 Task Settings

We input the answers to a short-answer question converted into text data using the automated handwriting recognition, and we output the corresponding predicted score. Subsequently, we demonstrate that our scoring model can predict the scores correctly by comparing the manual scores based on the rubric or scoring criteria. Regarding all the questions, we applied a single scoring model.

Sentence

"...Consider a situation where you travel to a country where you do not speak their language at all, and you have to ask for something. Regarding this case, **pointing will work like magic**...."

Question

What does the underlined phrase, **"pointing will work like magic"** indicate?

Student answer

It indicates that you can communicate your intentions, even if you do not speak the language.

Score: 3/3

Fig. 1. Example of a short-answer question conducted in 2018. It is originally written in Japanese and has been translated into English for reader's understanding.

Figure 2 shows the task flow. We evaluate the performance using the score, without modifying the character answer data and without adding any annotation to the answer. The part that should be correctly identified as "“や”" was identified as "“に”". The quality of the written letters was sometimes insufficient. This is because of stains that remained in the paper.

3.2 Handwriting Recognition

We employ the extracting, transforming, and loading (ETL) database, which has offline Japanese handwritten single characters. This database consists of nine datasets collected under different conditions [18]. Because the collected samples are written in separate boxes similar to the answer sheet of university entrance exams, the ETL database is appropriate for building an offline Japanese handwriting recognizer. This database covers the most common Japanese characters belonging to 2965 kanji (Japanese Industrial Standards : JIS Level 1) and 94 kana categories. Although there are more kanji categories, the characters of JIS Level 1 are mostly used daily and in examinations, whereas other kanji characters are rarely used.

Based on the success of the ensemble convolutional neural networks (CNNs) for Japanese historical character recognition [16], we also used an ensemble of multiple well-known CNN models. Our recognizer consists of a visual geometric group (VGG), MobileNet, residual network (ResNet), and ResNext networks with 16, 24, 34, and 50 layers, respectively [9,10,20,23].

To train these CNNs, we applied multiple transformations such as rotating, shearing, scaling, blurring, contrasting, and noise addition to avoid over-fitting problems because the database had only approximately one million samples in total. After training these CNNs using the ETL database, we fine-tuned them using 100 manually labeled samples from our collected Japanese handwritten answer database.

A trained neural network provides a prediction output as a k-dimensional vector of probabilities, where k is the number of categories for each character sample. These prediction outputs are averaged together with an equal weight of 1.0 to form an ensemble prediction output. Thus, the top-most prediction is the category with the highest probability in the ensemble prediction output. Figure 3 shows the procedure in which the CNN using 16, 24, and 50 layers is judged as "指." Here, the CNN using 34 layers is judged as "提," and finally it is judged correctly as "指."

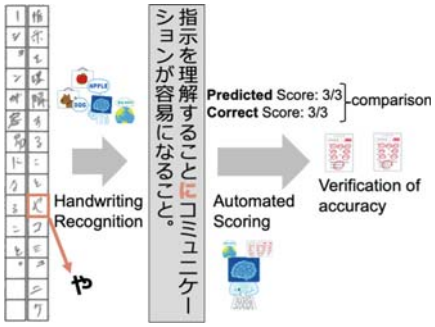

Fig. 2. Task flow

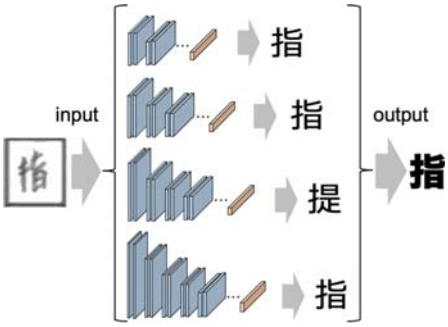

Fig. 3. Ensemble CNN handwriting recognition

Owing to the ambiguities of some characters, we also use an N-gram language model to correct misrecognized characters using the linguistic context. Considering every character of a text line, we computed the combined score based on the recognition and language scores of each character. First, the recognition score is the probability product of the previously recognized characters produced by the ensemble CNN recognizer. Second, the language score is the probability product of previous characters based on a five-gram Japanese language model that has been pre-trained by the Japanese Wikipedia corpus. Although N-grams are simple, they are sufficiently effective. Third, the combined score is a linear combination of the recognition and language scores with a weight of $\alpha \in [0, 1]$. Based on the combined score, we employ the beam search algorithm along the text line with a beam width of ten to export the top-ten candidates with the highest combined scores. However, only the highest combined score candidate was used for scoring in this experiment.

3.3 Scoring Procedure

The methods by [8] and [15], which perform the same type of SAS in Japanese, use an attention mechanism added to the bidirectional long-short term memory (Bi-LSTM). Their method outputs a predicted score based on each scoring criterion or rubric. However, our method does not accumulate scores for each scoring

criterion. It predicts the overall score. We explicitly utilize a multi-label classification model by fine-tuning it with Bidirectional Encoder Representations from Transformers (BERT) [4], which is pre-trained on Japanese Wikipedia.[1] If we consider the operation in large-scale tests, the scoring model should be implemented more efficiently. Nevertheless, we must utilize a better language model that is as accurate as possible.

The procedure is as follows (Fig. 4):

1. $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ is input as the written answer converted to text data using handwriting recognition, and the predicted score $s \in C = \{0, \ldots, N\}$ for the answer is provided as the output of the label.
2. The sentence $\boldsymbol{x}$ of the written answer is decomposed for each token, and a special token known as [CLS] is provided at the beginning of the sentence.
3. These token IDs are provided, and they are entered into the pre-trained BERT using the Japanese Wikipedia. Thereafter, we converted them into series of 768-dimensional vectors.
4. Whereas BERT is composed of all 12 layers, we concatenate the vectors of the [CLS] tokens of the last four layers of the hidden layer. Considering [4], combining them improved the document classification accuracy, compared to using only the [CLS] token vector in the final layer. Adam was used to optimize the model. The batch size was 16, and the number of epochs was five.
5. The vector of the combined classification [CLS] tokens is input into the classifier, and the predicted score s is output.

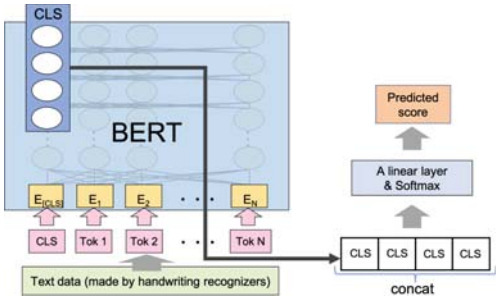

Fig. 4. Short-answer scoring model

3.4 Evaluation

The quadratic weighted kappa (QWK) [3] is often used as an evaluation index in SAS, and we used it in this study. The QWK is used for multilabel classification when an order relationship exists between labels. This index shows how well

[1] https://github.com/huggingface/transformers.

the correct and prediction labels match. The higher the value is, the better the prediction.

The QWK is calculated as:

$$\kappa = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}}, \tag{1}$$

where i and j represent the correct and predicted labels, respectively. $\boldsymbol{O}$ represents the ratio of each cell to the labels in the confusion matrix composed of the correct and predicted labels. $\boldsymbol{E}$ represents the expected value of the label belonging to each cell of the confusion matrix, assuming that the predicted and correct labels are independent.

$\boldsymbol{W}$ represents the penalty when the prediction is incorrect, and it is expressed as follows:

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2}, \tag{2}$$

where N represents the number of label classifications. $\boldsymbol{W}$ increases because the difference between the correct and predicted labels increases.

The QWK score is a ratio that can consider a value between -1 and 1. A negative QWK score indicates that the model is "worse than random." A random model should provide a score close to zero. Finally, the perfect predictions yielded a score of one. According to [14], Cohen suggested a kappa result of 0.81–1.00, which is interpreted as an approximately perfect agreement.

4 Experiments

4.1 Question Data

Six questions, including three questions each in 2017 and 2018, were classified based on these conditions and classification methods. The number of answers processed was approximately 60,000 in both 2017 and 2018. Table 1 shows the statistics of the scoring for each question. The question ID, number of answers, number of scoring conditions, score range, mean of the scores, standard deviation of the scores, and number of characters allowed are presented chronologically. We divided the data used for the BERT into 3:1:1 (= 60%:20%:20%) as the training, development, and evaluation sets. The scoring accuracy was evaluated using the QWK.

Table 1. Descriptive statistics on scoring for each question

Questions	# of answers	# of scoring conditions	Score range	Mean	# of characters allowed
2017 #Q1	62,222	4	0–6	4.46 ± 1.67	<50
2017 #Q2	61,777	3	0–2	1.51 ± 0.86	<25
2017 #Q3	59,791	4	0–5	0.43 ± 1.10	80–120
2018 #Q1	67,332	3	0–3	2.51 ± 0.88	<30
2018 #Q2	66,246	3	0–3	1.87 ± 1.14	<40
2018 #Q3	58,159	5	0–3	0.76 ± 1.07	80–120

4.2 Evaluation Results

Considering this experiment, the number of answer characters required at the university entrance level is relatively large, and the content is not plain. Regarding such cases, it is essential to know how large a sample is needed to guarantee the accuracy of the estimation.

Therefore, the sample size was changed to 50,000, 10,000, 5,000, 1,000, and 500, and the change in the QWK was observed. Table 2 shows the results, including the full-size data of approximately 60,000. The bold text indicates the best values.

Table 2. QWK for scoring each question

Questions	Sample size					
	Full size	50,000	10,000	5,000	1,000	500
2017 #Q1	0.978	**0.979**	0.967	0.946	0.883	0.679
2017 #Q2	**0.963**	0.949	0.934	0.922	0.818	0.884
2017 #Q3	**0.866**	0.836	0.705	0.680	0.473	0.276
2018 #Q1	**0.976**	0.968	0.974	0.914	0.863	0.820
2018 #Q2	**0.954**	0.945	0.923	0.903	0.796	0.724
2018 #Q3	**0.944**	0.929	0.916	0.894	0.783	0.753

The following can be obtained from the steps above.

1. We observe that the accuracy is kept high by the method for all six questions, regardless of the type of question. Even in the worst case of Q3 in 2017, the QWK is 0.86; otherwise, it is 0.94 or higher.
2. Essentially, the larger the sample size is, the better the accuracy. This indicates that the accuracy does not converge, which is an unexpected result. The sample size of 60,000 seems large enough in a typical test. Nevertheless, it shows that a more significant number is needed to improve the accuracy of the prediction.

This indicates in a sentence of a certain length, the variation in expressions is highly diverse. Because the number of characters increased, the number of variations increased exponentially, even if we have sufficient answer patterns that would not be significant. Therefore, the learning never converges.

3. The easier the question is, the higher the scoring rate, and the better the estimation accuracy. In both 2017 and 2018, Q1 was the easiest, and Q3 was the most difficult. The accuracy of Q1 was higher than that of Q3. This tendency did not depend on the number of scores.

5 Ablation Study

We observed the effect on scoring accuracy in our model from two perspectives. First, we considered the accuracy of handwriting recognition. We examined how the recognition rate affected the overall scoring accuracy. Second, we considered the position of the layer in the language-processing model. We changed the information position extracted from the 12 layers of the BERT model and verified how the change affected the overall scoring accuracy.

5.1 Effect of the Handwriting Recognition Models Used

To investigate the effect of the handwritten character recognition part on the scoring accuracy, we compare the original ensemble model of four methods with other methods. The compared methods are as follows:

1. No language model: This is a character recognition model without correction of misrecognized characters by the N-gram language models.
2. VGG only: This is a single character recognition model without ensemble learning.
3. DenseNet only: This is also a single character recognition model without ensemble learning.
4. Ensemble 5: This is a character recognition model with ensemble learning of five character recognition models.

Table 3 compared the QWK using each of the output results.

Table 3. Comparison of QWK by five methods

Questions	The handwriting recognition models				
	Original	No language model	VGG	DenseNet	Ensemble5
2017#Q1	0.978	0.975	0.977	0.974	**0.980**
2017#Q2	**0.963**	0.957	0.957	0.952	0.959
2017#Q3	**0.866**	0.847	0.844	0.820	0.830
2018#Q1	**0.976**	0.973	0.972	0.970	0.970
2018#Q2	**0.954**	0.950	0.952	0.953	0.953
2018#Q3	**0.944**	0.937	0.933	0.935	0.941

This shows that the model with ensemble learning with multiple character recognition models has a higher overall accuracy than the model with a single character recognition model. In addition, the results show that the accuracy of the models with the modification in the language model is higher than that of the models without modification. Moreover, increasing the number of ensemble learning models did not change significantly, considering the accuracy. Considering these results, we observed that the overall accuracy was affected by both the language model changes and character recognition model quality. Moreover, we found that the overall accuracy was limited by improving the quality of the character recognition model.

5.2 Effect of the Information Retrieved from the BERT Model

We investigated the effect of the different linguistic information retrieved from BERT on the scoring accuracy. The BERT used in our study consists of 12 layers, and each layer is known to contain different information [11]. Specifically, the layers close to the input, middle, and output parts possess morphological information, syntactic information, and information that focuses on the semantic information, respectively. We divided the BERT model into three parts: a layer near the input, a middle part, and a layer near the output. Thereafter, we examined the differences in the scoring accuracy between the three parts. Layers 1–4, 5–8, and 9–12 were extracted from the input section. The output from each layer was input into the linear layer, and the score was predicted. Table 4 lists the results for each accuracy. We observed that the scoring accuracy was the highest when the information of layers 9–12 was extracted for each problem.

Table 4. Comparison of QWK by the different extraction layers

Questions	The part of layers		
	1–4	5–8	9–12
2017#Q1	0.977	0.977	**0.978**
2017#Q2	0.952	0.955	**0.963**
2017#Q3	0.830	0.832	**0.866**
2018#Q1	0.969	0.972	**0.976**
2018#Q2	0.951	0.950	**0.954**
2018#Q3	0.936	0.939	**0.944**

This indicates that the system is paying particular attention to the semantic information when performing automatic scoring tasks. Particularly, QWK in 2017#Q3 was different by 3.0 or more among all the questions, and the difference was outstanding.

6 Summary and Conclusions

We have investigated a fully automated scoring method for short-answers using handwriting recognition data and have evaluated the system's performance using a large-scale national test. "Fully" indicates that there is no need to annotate the scoring data or convert the handwritten text manually. We used very large data conducted in two trial tests for university common entrance examinations and used a pre-trained BERT model for scoring. We made the following observations.

1. When the data is sufficiently large, our method increases the scoring accuracy without annotation and converts the handwritten text manually.
2. When we consider 25 to 120 character answers, learning often does not converge, even with a data size of 50,000.
3. Even if some errors are caused by handwriting recognition, the accuracy of scoring is guaranteed to some extent using the current technology.

This study reports the actual accuracy at the current technical level in a procedure without human intervention. Despite the variety in the types of questions we considered, such as the number of characters in the answer and the difficulty level, we could predict the scores with high accuracy in all cases. This suggests that our procedure is effective for all short-answer questions, and SAS is suitable for large-scale testing using the current technology. In addition, our study demonstrates the usefulness of the method for utilizing handwritten character recognition models in SAS. We can serve as an opportunity to develop a new learning method for educational application settings, where students often use handwriting.

Acknowledgements. This work was supported by JSPS KAKENHI Grant Number JP20H04300 and JST A-STEP Grant Number JPMJTM20ML.

References

1. Alikaniotis, D., Yannakoudakis, H., Rei, M.: Automatic text scoring using neural networks. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 715–725. Association for Computational Linguistics, Berlin (2016). https://doi.org/10.18653/v1/P16-1068
2. Burstein, J., Tetreault, J., Madnani, N.: The e-rater automated essay scoring system. In: Shermis, M.D., Burstein, J. (eds.) Handbook of Automated Essay Evaluation, Chap. 4, pp. 55–67. Edwards Brothers Inc., New York (2013)
3. Cohen, J.: Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. **7**(4), 213–220 (1968)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis (2019). https://doi.org/10.18653/v1/N19-1423

5. Dong, F., Zhang, Y.: Automatic features for essay scoring - an empirical study. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1072–1077. Association for Computational Linguistics, Austin (2016). https://doi.org/10.18653/v1/D16-1115
6. Drid, T.: The fundamentals of assessing EFL writing. Psychol. Educ. Stud. **11**(1), 292–305 (2018). https://doi.org/10.35156/1192-011-001-017 https://doi.org/10.35156/1192-011-001-017 https://doi.org/10.35156/1192-011-001-017 https://doi.org/10.35156/1192-011-001-017
7. Foltz, P.W., Streeter, L.A., Lochbaum, K.E., Landauer, T.K.: Implementation and applications of the intelligent essay assessor. In: Shermis, M.D., Burstein, J. (eds.) Handbook of Automated Essay Evaluation, Chap. 5, pp. 55–67. Edwards Brothers Inc., New York (2013)
8. Funayama, H., et al.: Preventing critical scoring errors in short answer scoring with confidence estimation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 237–243. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-srw.32
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
11. Jawahar, G., Sagot, B., Seddah, D.: What does BERT learn about the structure of language? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3651–3657. Association for Computational Linguistics, Florence (2019). https://doi.org/10.18653/v1/P19-1356
12. Li, Z., Tomar, Y., Passonneau, R.J.: A semantic feature-wise transformation relation network for automatic short answer grading. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6030–6040. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021)
13. Lottridge, S., Wood, S., Shaw, D.: The effectiveness of machine score-ability ratings in predicting automated scoring performance. Appl. Measur. Educ. **31**(3), 215–232 (2018)
14. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochem. Med. (Zagreb) **22**(3), 276–282 (2012)
15. Mizumoto, T., et al.: Analytic score prediction and justification identification in automated short answer scoring. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 316–325 (2019)
16. Nguyen, H.T., Ly, N.T., Nguyen, K.C., Nguyen, C.T., Nakagawa, M.: Attempts to recognize anomalously deformed kana in Japanese historical documents. In: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing, pp. 31–36. HIP2017, Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3151509.3151514
17. Riordan, B., Horbach, A., Cahill, A., Zesch, T., Lee, C.M.: Investigating neural architectures for short answer scoring. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 159–168. Association for Computational Linguistics, Copenhagen (2017). https://doi.org/10.18653/v1/W17-5017

18. Saito, T., Yamada, H., Yamamoto, K.: On the database ETL9 of handprinted characters in JIS Chinese characters and its analysis. Trans. IECE Jpn. **J68-D**(4), 757–764 (1985)
19. Schultz, M.T.: The intellimetric automated essay scoring engine - a review and an application to chinese essay scoring. In: Shermis, M.D., Burstein, J. (eds.) Handbook of Automated Essay Evaluation, Chap. 6, pp. 55–67. Edwards Brothers Inc, New York (2013)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. 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
22. Uto, M., Okano, M.: Robust neural automated essay scoring using item response theory. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12163, pp. 549–561. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52237-7_44
23. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)