Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1250–1257
Marseille, 11–16 May 2020

Automated Essay Scoring System for Nonnative Japanese Learners

Reo Hirao, Mio Arai, Hiroki Shimanaka, Satoru Katsumata, Mamoru Komachi
Tokyo Metropolitan University
6-6 Asahigaoka, Hino-city, Tokyo, Japan
{hirao-reo@ed., arai-mio@ed., shimanaka-hiroki@ed., katsumata-satoru@ed., komachi@}tmu.ac.jp

Abstract

In this study, we created an automated essay scoring (AES) system for nonnative Japanese learners using an essay dataset with annotations for a holistic score and multiple trait scores, including content, organization, and language scores. In particular, we developed AES systems using two different approaches: a feature-based approach and a neural-network-based approach. In the former approach, we used Japanese-specific linguistic features, including character-type features such as "kanji" and "hiragana." In the latter approach, we used two models: a long short-term memory (LSTM) model (Hochreiter and Schmidhuber, 1997) and a bidirectional encoder representations from transformers (BERT) model (Devlin et al., 2019), which achieved the highest accuracy in various natural language processing tasks in 2018. Overall, the BERT model achieved the best root mean squared error and quadratic weighted kappa scores. In addition, we analyzed the robustness of the outputs of the BERT model. We have released and shared this system to facilitate further research on AES for Japanese as a second language learners.

Keywords: Learner Written Essay, Japanese, Automated Essay Scoring

1. Introduction

Automated essay scoring (AES) is a task in which computer technology is used to evaluate written text. Humans find it difficult to evaluate a large number of essays. In this light, AES has emerged as one of the most important educational applications of natural language processing. The major weakness of existing scoring systems is that they only provide a single holistic score that summarizes the quality of an essay. Such a score provides little feedback, especially for a language learner. For example, when a system only returns a low holistic score, the learner cannot understand which aspect of the essay is inadequate without language teachers.

To address this problem, some studies scored various dimensions of essay quality, such as prompt adherence (Persing and Ng, 2014), organization (Persing et al., 2010), and coherence (Miltsakaki and Kukich, 2004). Furthermore, some datasets are available for evaluating additional dimensions of essay quality in English (Mathias and Bhattacharyya, 2018). However, only a few evaluation datasets are available for Japanese writings, and even fewer Japanese learner essay datasets are. Nevertheless, Tanaka and Kubota overcame this lack of availability of Japanese learner datasets by creating a new dataset for Japanese learners (Tanaka and Kubota, 2016). In this dataset, they annotated a holistic score and three trait scores—content, organization, and language scores—of essay quality on a learner's ability.

With the Japanese learner corpora annotated for multiple traits, we created essay scoring systems for Japanese learners using two different machine learning approaches: feature-based models and neural-network-based models. We created the feature-based models using linguistic features based on (Lee and Hasebe, 2017). However, whether these features are enough to perform AES is unclear. To make a model without developing features, we created a neural-network-based model using long short-term memory (LSTM) units (Hochreiter and Schmidhuber, 1997) and an attention mechanism (Bahdanau et al., 2015). We also created another neural-network-based model using bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019).

The main contributions of this study are as follows:

- The first AES system for Japanese learners is created to evaluate essays in both a holistic way and using multiple traits.
- The first neural-network-based Japanese AES system is created. Furthermore, the system outputs for essays, including an illegal input essay, were analyzed using both the neural-network-based and feature-based systems.

We have released and shared this system to facilitate further research on AES[1].

2. Related Work

In recent years, many AES engines have been developed (Page, 1966; Shermis and Burstein, 2013). Some of them provide not only a holistic score but also other scores of essay quality. In 2012, Kaggle organized an AES competition called Automated Student Assessment Prize (ASAP)[2]. The dataset of this competition contains more than 10,000 essays with holistic scores, and it is now used in English AES research. However, few such Japanese datasets are available because of insufficient resources and the difficulty of obtaining essay data.

The Japanese essay scoring system (Jess) (Ishioka and Kameda, 2006) has been created for scoring essays in college-entrance exams. Jess can provide not only a holistic score but also multiple trait scores—rhetoric, organization, and content scores—using statistical methods. Jess uses Mainichi Daily News data to measure the difference between an input essay and expert-written essays using linguistic features such as sentence length and

[1] The systems are available at https://github.com/reo11/aes-for-japanese-learner

[2] https://www.kaggle.com/c/asap-aes

Data Name	No. of Essays with Holistic Score	No. of Essays with Multiple-trait Scores	No. of Learner Mother Tongues	Prompt
I-JAS	578	56	12	"Fast Food and Home Cooking"
TK	212	32	2	"Fast Food and Home Cooking" or "Traditional Education and E-Learning"
EU	68	60	11	"Individual Trip and Package Trip" or "Dining Out and Self-Catering"

Table 1: Details of the GoodWriting dataset. The Prompt column shows given topics for the essays. In TK data and EU data, the essay is written on either one of the two prompts. As shown in the Prompt column, all essays in the dataset are argumentative essays.

ID	A	B	C	D	E	F	ave.
κ	0.57	0.61	0.59	0.45	0.60	0.53	0.56

Table 2: Kappa coefficient between each annotator and final scores in the holistic evaluation.

"kanji"/"kana" ratio.

JWriter (Lee and Hasebe, 2017) is an AES system for language learners that provides a holistic score from one to three points. This system provides scores via linear regression using linguistic features such as the average number of words, number of part-of-speech (POS) tags, and total number of characters. In addition, it can predict a holistic score robustly with just a small training dataset; however, because it does not use surface information, it cannot predict multiple traits that require surface text information. Moreover, if the regression equation is known to the learners, they could cheat the system easily.

By contrast, neural-network-based methods do not need to create features and have produced state-of-the-art results in various datasets (Ke and Ng, 2019). As a result, the neural approach for AES has been actively studied in recent years (Taghipour and Ng, 2016). However, no neural-network-based AES system is available for the Japanese language; furthermore, the BERT model has not been applied for an AES task with multiple dimensions thus far. Therefore, we create such a system and report the obtained results. Further, because the features in the neural approach are not explicit to the learners, we analyze how robust neural-network-based models are to cheating.

3. Dataset

We used the GoodWriting dataset[3]. This dataset contains more than 800 essays written by Japanese learners overseas. Each essay was annotated by three annotators, and the final score is determined by the middle value of the three scores. Figure 1 shows the distribution of the score of each essay in the dataset. Holistic scores were given for all 858 essays, and three trait scores—content, organization, and language scores—were given for 148 essays. The holistic scores and three trait scores are scored from one to six points in the dataset.

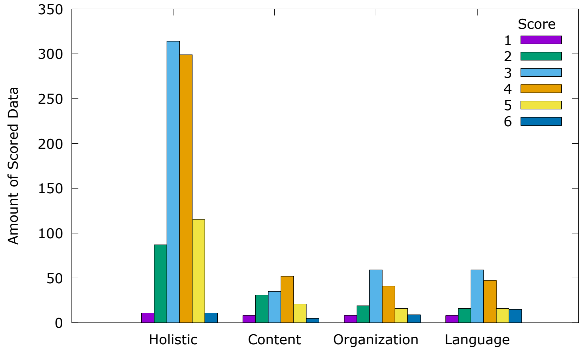


Figure 1: Distribution of the GoodWriting dataset. The bar graph in the figure represents the number of essays with scores of 1, 2, 3, 4, 5, and 6 from left to right.

The GoodWriting dataset comprises three different types of data[4]: I-JAS data[5], TK data, and EU data. Table 1 shows details about the GoodWriting dataset. I-JAS data comprises composition data of Japanese language learners from 12 countries. TK data comprises data of Japanese learners whose native language is English or Chinese (Tanaka and Kubota, 2016). EU data comprises data of Japanese learners from 11 countries.

3.1. Traits Evaluated in Essays

The dataset contains a holistic score and three trait scores—content, organization, and language scores—of essay quality. For creating the dataset, the annotators used an evaluation flowchart[6] to minimize the differences between annotators (Tanaka et al., 2009).

Holistic The holistic evaluation focuses on the overall impression without being constrained by details. The main consideration in holistic evaluation is whether the problem has been achieved. Achieving a problem means that the author is giving a comparative opinion. Further, if the content, organization, and language levels exceed a certain level, the score will increase.

[3] https://goodwriting.jp/wp/?lang=en

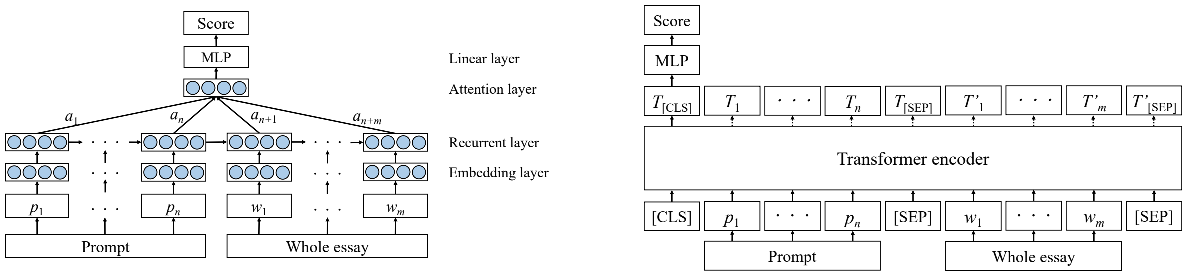
[4] https://goodwriting.jp/wp/system-data

[5] http://lsaj.ninjal.ac.jp/?cat=3

[6] The evaluation flowchart of each trait is available here: https://goodwriting.jp/wp/system-flowcharts

Target trait	Feature
Holistic	Total numbers of characters, morphemes, commas, sentences, and paragraphs.
Content	Total numbers of morphemes that appear in common with the essay and the prompt.
Organization	Average numbers of characters, morphemes, and commas in each paragraph. Ratio of characters, morphemes, and sentences in the first paragraph to those in the entire essay. Ratio of characters, morphemes, and sentences in the last paragraph to those in the entire essay.
Language	Ratio of each POS tag. Ratio of “hiragana,” “katakana,” and “kanji.”

Table 3: Features used in our feature-based models. The target trait column shows which label the features were designed for. Note that all the features were used to train the feature-based models.



(a) AES system with unidirectional LSTM and the attention mechanism. The weights of the attention mechanism are represented by a in the figure.

(b) AES system with BERT. The output embeddings of the transformer encoder are represented by T and T' in the figure.

Figure 2: Two neural-network-based models. n is the number of words in the prompt and m is the number of words in the essay. Each word in the prompt is represented by p and each word in the essay is represented by w.

Content The content evaluation measures essays from two perspectives: objective and content. The objective perspective involves evaluating the task achievement, that is, whether a comparison is made and opinions are stated. In this dataset, the validity of the comparison is evaluated. The content perspective involves evaluating the consistency of the main idea and the validity of the supporting information.

Organization The organization evaluation measures essays from three perspectives: compositional awareness (awareness to write considering the order of stories), paragraph awareness (awareness to write one set of things in one paragraph), and macrostructure (introduction, main argument, and conclusion).

Language The language evaluation measures essays from three perspectives: readability, adequacy, and diversity. The minimum level is that people who are not Japanese teachers can barely read by an educated guess. From there, the score increases as each perspective level exceeds a certain level.

3.2. Details about Annotation

The essays were annotated by six annotators. Specifically, each essay was annotated by three out of these six annotators. The final score of the essay is the middle value of the three scores. All annotators are Japanese native speakers who have some experience in evaluating essays. We evaluated each annotator’s score by calculating Cohen’s kappa coefficient between the annotated scores and the final scores. Table 2 lists the kappa coefficient between each annotator’s scores as well as the final scores of the holistic evaluation. The average of all six annotators’ kappa coefficients is 0.56. According to Landis and Koch (1977), a kappa coefficient of 0.41–0.60 indicates moderate agreement. Although the scoring criteria were applied as per the evaluation flowchart, some of the annotators were severe; therefore, the resulting agreement scores were moderate. In particular, between the most severe and the gentlest of the six annotators, there was an average difference of about 0.5 points on a one to six–point rating.

4. AES Systems

The AES task using an annotated dataset with an integer number can be recasted as (1) a regression task with the goal of predicting an integer score with a real number and (2) a classification task with the goal of classifying an essay to one of the classes corresponding to the scores.

We created a Japanese AES system using regression models for two reasons. First, state-of-the-art methods for various AES datasets use a regression approach (Ke and Ng, 2019). Second, even if the amount of data for each score is biased, it does not overfit as much as the classification models do. We propose five models using the GoodWriting dataset: three models with a feature-based method and two models with a neural method. Each model is described in the following subsections.

4.1. Feature-based Methods

We follow the feature set described by Zesch et al. (2015). Additionally, we use linguistic features unique to Japanese, including character-type features such as "kanji" and "hiragana," that are used in jWriter (Lee and Hasebe, 2017) and GoodWriting Rater[7]. Table 3 shows the features used in our systems. Each feature is designed for either holistic, content, organization, or language traits. All the features listed in Table 3 are used when predicting individual scores. We apply linear regression and linear support vector regression (SVR), which have been used in recent studies (Ke and Ng, 2019). We also used random forest regression to examine a decision-tree approach, unlike previous studies.

4.2. Neural Methods

We present two different neural-network-based models: LSTM model and BERT model. Figure 2-(a) shows the architecture of the LSTM model. In our LSTM model, we predict the essay score in five steps. First, morphological analysis is used to divide sentences into morphemes because there is no delimiter for separating words in Japanese. Second, each morpheme is converted into a vector. Third, each vector is inputted into a recurrent neural network (unidirectional LSTM) across sentences. The whole vectors in the essay are taken as a series in this model. Fourth, the hidden layer vectors are aggregated using an attention mechanism. Finally, the vector is converted to a single scalar in the linear layer.

BERT is a fine-tuning-based language representation model that uses a transformer architecture (Vaswani et al., 2017) trained on two tasks: masked language model and next sentence prediction. Recently, Nadeem et al. (2019) applied BERT to AES systems. Figure 2-(b) shows the architecture of our BERT model. The BERT model takes a tokenized prompt and a whole essay as an input. When tokenizing the input, a [CLS] token is added at the beginning. The prompt and the essay are distinguished by a [SEP] token added at the end. If the essay length exceeds the upper limit of the sequence, the rest of the essay will be excluded so that it will fit in the sequence including three tokens: a [CLS] token and two [SEP] tokens. The prompt and the essay sequences are inputted into the transformer encoder and become hidden layer sequences. The hidden layer corresponding to the [CLS] token ($T_{[\mathrm{CLS}]}$) represents the distributed representation of a prompt and an essay. The final numerical value is obtained by a multilayer perceptron (MLP) using this distributed representation. When training the MLP, the transformer encoder is also retrained.

5. Experiments

5.1. Evaluation Metrics

In our AES system, the output score of each essay is a real number. We use two evaluation metrics to evaluate our models. The first metric is the root mean squared error (RMSE). In this study, every model predicts a real number. Thus, we used the RMSE score to measure the difference between the gold scores and the predicted scores. We also used the RMSE score to train the models.

RMSE is calculated using Equation (1):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{1}$$

where N is the total number of essays, y_i is the actual score of i-th essay, and $\hat{y}_i$ is the predicted score of i- th essay.

The second metric is a quadratic weighted kappa (QWK). The QWK was used as an evaluation metric in the ASAP competition and in recent studies (Taghipour and Ng, 2016; Cozma et al., 2018). This metric gives a square penalty depending on the distance between integer values: gold scores and predicted scores.

QWK is calculated using Equation (2):

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \tag{2}$$

where matrix O is calculated such that $O_{i,j}$ is the number of essays that was scored i by the human annotator and scored j by the AES system. Matrix E is calculated as the outer product between the histogram vectors of the actual and predicted scores. The matrices E and O are normalized such that E and O have the same sum. Weight $w_{i,j}$ is calculated using Equation (3):

$$w_{i,j} = \frac{(i-j)^2}{(R-1)^2} \tag{3}$$

where R is the maximum score of the essay.

5.2. Setting

We use MeCab (ver. 0.996)[8] as a morphological analysis tool when we create the features listed in Table 3. In the feature-based methods, we use the IPA dictionary (ver. 2.7.0) for the MeCab dictionary. We use optuna[9] to optimize hyperparameters in feature-based models.

The word embeddings in the LSTM model are generated using pre-trained word2vec[10]. UniDic (Den et al., 2008) is used as a MeCab dictionary because the vocabulary in the pre-trained word2vec was taken from UniDic. UniDic and IPADic are both dictionaries developed for Japanese morphological analysis. UniDic segments morphemes into shorter units, called "short units," while IPADic does not include these. Because morphemes are divided into short units, there is little ambiguity when manually creating a dictionary, and the dictionary is robust against colloquial sentences. Our LSTM model has several hyperparameters that need to be set. The dimension of the word embedding is 300, and the dimension of the hidden layer is 512. We use a unidirectional LSTM unit in the recurrent layer and two dense layers in the linear layer, one for reducing the vector's dimension and the other for providing a score. It uses the mean squared error as a loss function, a rectified linear unit as an activation function, and early stopping with patience = 20. We regularize the network by using dropout,

[7] https://goodwriting.jp/wp/system-ml/

[8] https://taku910.github.io/mecab

[9] https://github.com/pfnet/optuna

[10] https://github.com/Kyubyong/wordvectors

Metric	Model	Holistic	Content	Organization	Language	Average
RMSE	Linear SVR	1.016 ± .0449	1.222 ± .0704	0.960 ± .0238	1.005 ± .0160	1.051 ± .0480
	Linear Regression	0.771 ± .0007	1.246 ± .0246	1.012 ± .0197	1.072 ± .0290	1.025 ± .0471
	Random Forest	0.769 ± .0007	1.052 ± .0110	0.924 ± .0075	1.048 ± .0103	0.948 ± .0206
	LSTM	0.948 ± .0044	1.135 ± .0195	1.082 ± .0092	1.227 ± .0133	1.098 ± .0216
	BERT	**0.714 ± .0009**	**0.993 ± .0119**	**0.902 ± .0111**	**0.898 ± .0086**	**0.876 ± .0182**
QWK	Linear SVR	0.482 ± .0055	0.409 ± .0223	**0.602 ± .0066**	0.597 ± .0130	0.523 ± .0181
	Linear Regression	0.533 ± .0010	0.358 ± .0239	0.552 ± .0148	0.571 ± .0181	0.503 ± .0213
	Random Forest	0.519 ± .0024	0.400 ± .0197	0.528 ± .0106	0.462 ± .0138	0.477 ± .0140
	LSTM	0.247 ± .0134	0.362 ± .0250	0.441 ± .0121	0.332 ± .0138	0.346 ± .0204
	BERT	**0.621 ± .0018**	**0.494 ± .0189**	0.540 ± .0131	**0.621 ± .0094**	**0.569 ± .0135**

Table 4: RMSE and QWK scores of AES models. The best score in each column is indicated in bold. The average score is the mean value of the holistic, content, organization, and language scores.

and we set the dropout probability to 0.5.

We use the BERT$_{\text{base}}$ architecture (Devlin et al., 2019) in our experiments. We use a pre-trained BERT model with SentencePiece for Japanese text[11] of which the vocabulary is obtained by SentencePiece (Kudo and Richardson, 2018). We selected this model because it can input the longest sequence (512 tokens) among the pre-trained Japanese models that are publicly available. The vocabulary size of this model is 32,000. We fine-tuned the pre-trained BERT model with learning rate of 5e-5, batch size of 4, and maximum sequence length of 512. We set the dropout probability for all fully connected layers in the embeddings, encoder, and pooling layers to 0.1 and the dropout ratio for the attention probabilities to 0.1.

All scores are evaluated as five-fold cross-validation. Additionally, because the evaluation results vary depending on the seed value of data division, the experiment was performed with five seeds, and the final score is calculated as the mean of the five scores.

5.3. Result

We compared the scores of each model and analyzed the results. Table 4 shows the RMSE and QWK scores of each model. The average column shows the mean holistic, content, organization, and language scores. The results indicate that the BERT model has the highest holistic and trait (except for organization) scores for both RMSE and QWK.

A comparison of the holistic evaluation with the evaluation of other traits indicates that the former is up to 0.12 higher than the other traits for QWK. This may be because the holistic evaluation was performed for more than 800 essays and the other evaluations, for only about 150 essays.

5.4. Analysis

Feature-based methods Table 4 shows that the linear SVR model had the highest QWK score and the lowest variance in the organization trait score. Further, the other feature-based methods also had higher QWK scores in the trait than BERT, which had the highest overall average score. To determine why the feature-based methods had effective scores in this trait compared to those of BERT, we analyzed the weights of the features of linear SVR for holistic scores and the organization trait score.

Tables 5 and 6 show the top ten weights assigned to the features when the linear SVR predicted the holistic score and the organization trait scores. The weights are calculated using the average of the features in the five-fold cross-validation process. Intuitively, "Number of common morphemes in the prompt and the essay" appears in holistic features. In other words, a story related to the prompt is more highly rated than one not related to the prompt. Most top ten features were related to the essay length, such as "Number of different morphemes," "Number of characters," and "Number of sentences." In AES, there is a high correlation between the essay length and the holistic score, and a long essay tends to get a high holistic score (Shermis and Burstein, 2003).

The most important weight for the organization trait is "Number of paragraphs." Essays with appropriate paragraphs are highly rated because the organization trait evaluates logical structures such as introductions, main articles, and conclusions. The ninth and tenth features of organization weights include the ratio of the first paragraph. This shows that the features that are related to the balance of the paragraph have high importance. The feature related to the number of paragraphs is not explicitly captured by a neural-network-based model unless explicitly inputted; thus, entering an accurate number as a feature is thought to improve the score of the organization trait.

In the original dataset, most essays had a length of around six paragraphs at most, but some essays contained more than ten paragraphs. Essays with more than ten paragraphs containing a few sentences each had low scores such as one or two. Unnecessary paragraph divisions are considered to adversely affect the composition score because they violate paragraph awareness and macrostructure.

Neural-network-based methods The BERT model used in this study pre-trains the Masked Language Model (MLM) using a large amount of sentence data. This enables the BERT model to score essays considering context in contrast to the LSTM model, and thus, the BERT model scored higher than the LSTM model in Table 4. The BERT model in this study provided better scores than other meth-

[11] https://github.com/yoheikikuta/bert-japanese

No.	Top ten features in holistic score prediction	Weight
1	Number of common morphemes in the prompt and essay.	0.0331
2	Number of different morphemes.	0.0261
3	Average number of commas in each sentence.	-0.0109
4	Number of characters.	0.0092
5	Average number of commas in each paragraph.	0.0088
6	Number of sentences.	0.0080
7	Average number of characters in each sentence.	-0.0053
8	Number of paragraphs.	-0.0033
9	Number of morphemes.	-0.0030
10	Average number of morphemes in each sentence.	-0.0030

Table 5: Top ten weights of linear SVR in holistic score prediction. The weights are sorted in descending order of absolute values.

No.	Top ten features in organization trait score prediction	Weight
1	Number of different morphemes.	0.0238
2	Number of common morphemes in the prompt and essay.	0.0205
3	Number of characters.	0.0141
4	Number of paragraphs.	-0.0128
5	Average number of characters in each sentence.	-0.0113
6	Number of morphemes.	-0.0029
7	Average number of morphemes in each sentence.	-0.0029
8	Average number of commas in each sentence.	-0.0029
9	Average number of morphemes in each paragraph.	-0.0022
10	Average number of commas in each paragraph.	-0.0017

Table 6: Top ten weights of linear SVR in organization trait score prediction. The weights are sorted in descending order of absolute values.

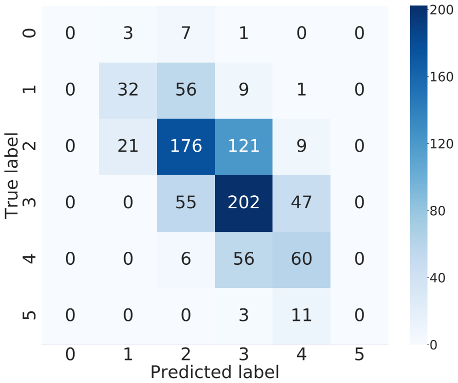


Figure 3: Confusion matrix of the holistic scores predicted by the BERT model and the actual holistic scores. The horizontal axis represents the essay score predicted by the system and the vertical axis, the actual essay score. The numbers in the matrix indicate the number of essays that correspond to the predicted scores and actual scores.

ods; however, this model has some disadvantages. Figure 3 shows a confusion matrix of the holistic scores predicted by the BERT model and the actual holistic scores. This figure shows that errors are within around ±1; however, the system does not predict any score of one or six points. The problem with the current BERT model is that it can only predict a number of points close to the average value, unlike the feature-based method. This is attributed to the fact that only some data have the highest/lowest scores and the systems are trained with a loss function of the mean squared error. This is because predicting the mean value will minimize the penalty of the mean squared error.

This problem could be solved using one of two approaches. The first approach is to develop a system that can flexibly predict scores with little data. The second approach is to create a new dataset. Because very few datasets are available in the AES field, especially fewer datasets annotated in multiple traits, we should focus on creating a new dataset.

Adversarial essay To check the scores that the AES system outputs for unseen data, we entered several types of essays: an essay with a high score, an essay with a low score, and an adversarial essay written with only one repeating character. Most existing AES systems may assign high scores to essays written with only one character or to essays generated with random words. To confirm the robustness of the proposed system against this problem, we prepared an adversarial essay.

Table 7 shows the result of the linear regression model and the BERT model using test inputs. The rows of essays A and B are the results of two real essays with high and

Essay	Type	Holistic	Cont.	Org.	Lang.
A	gold	5	5	5	5
	feature	5	**4**	5	**6**
	neural	5	5	5	5
B	gold	2	2	3	3
	feature	2	**1**	**1**	3
	neural	2	2	3	**2**
Adv.	gold	1	1	1	1
	feature	1	**6**	1	**3**
	neural	**2**	**2**	**2**	**2**

Table 7: Essay scores predicted via linear regression model and BERT model. The type column shows the type of score: gold score or predicted scores by feature-based/neural-network-based models. Essays A and B are real essays written by learners. Scores that differ from the gold score are indicated in bold. Essay Adv. stands for adversarial essay, which is written with only one character (e.g., “おおおおお、おおお...”.)

low scores. The feature-based model predicted a score that was two points lower than the actual content and organization trait scores in essay A and a score that was two points lower than the organization trait score in essay B. In contrast, the neural-network-based model predicted the score correctly for essay A and predicted a score that was only one point lower than the actual language trait score in essay B. These results indicate that the neural-network-based model makes fewer mistakes than the feature-based model, and the fluctuation range of the score is narrow even if a mistake is made.

In the adversarial essay, the feature-based model predicted extremely high or low scores. Further, this model may provide a high score for an unexpected input. In contrast, the neural-network-based model predicted low scores for all columns. Thus, the neural-network-based model is robust against an unexpected essay. The robustness of the BERT model can be attributed to the use of a pre-trained model. On the contrary, feature-based methods do not use pre-trained models; furthermore, LSTM models use only the pre-trained word2vec approach, while BERT trains the MLM with a large amount of sentence data. As mentioned earlier, because MLM can learn hidden layer information including the context, it is believed that an artificial essay, such as an adversarial essay, would be recognized as an invalid essay and would be awarded a low score.

6. Conclusion

In this study, we created AES systems for nonnative Japanese learners. We created these systems using a dataset with annotations on holistic scores and multiple trait scores—content, organization, and language scores. We compared each evaluation trait and the average score of all traits. The results indicated that the neural approach using BERT achieved the highest score among five models.

We confirmed the robustness of the BERT model with three essays: an essay with a high/low score and one written with only one character. As a result, we found that the BERT model is more robust against unexpected inputs than the feature-based models.

7. Acknowledgements

This research is partly supported by scientific research grant project (JSPS) no. 26284074. We thank Mari Tanaka (PI), Yukari Tsubone, Yoko Kageyama, Shin Abe, Aiko Sasaki, and Yo Usami for their help in creating the dataset. We would like to especially thank Shin Abe for providing the adversarial essay.

8. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, California, May. Conference Track Proceedings.

Cozma, M., Butnaru, A., and Ionescu, R. T. (2018). Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503–509, Melbourne, Australia, July. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ishioka, T. and Kameda, M. (2006). Automated Japanese essay scoring system based on articles written by experts. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 233–240, Sydney, Australia, July. Association for Computational Linguistics.

Ke, Z. and Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6300–6308, Macao, China, July. International Joint Conferences on Artificial Intelligence Organization.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Mathias, S. and Bhattacharyya, P. (2018). ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1169–1173, Miyazaki, Japan, May. European Language Resources Association.

Miltsakaki, E. and Kukich, K. (2004). Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55.

Nadeem, F., Nguyen, H., Liu, Y., and Ostendorf, M. (2019). Automated essay scoring with discourse-aware neural models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 484–493, Florence, Italy, August. Association for Computational Linguistics.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Persing, I. and Ng, V. (2014). Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1534–1543, Baltimore, Maryland, June. Association for Computational Linguistics.

Persing, I., Davis, A., and Ng, V. (2010). Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA, October. Association for Computational Linguistics.

Shermis, M. D. and Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Shermis, M. D. and Burstein, J. C. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Taghipour, K. and Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas, November. Association for Computational Linguistics.

Tanaka, M., Nagasaka, A., Narita, T., and Sugai, H. (2009). Writing assessment workshop for Japanese as a second language: Examining a scoring rubric. *Japanese-Language Education around the Globe*, 19:157–176.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, Long Beach, California, December. Curran Associates, Inc.

Zesch, T., Wojatzki, M., and Scholten-Akoun, D. (2015). Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–232, Denver, Colorado, June. Association for Computational Linguistics.

9. Language Resource References

Den, Y., Nakamura, J., Ogiso, T., and Ogura, H. (2008). A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1019–1024, Marrakech, Morocco, May. European Language Resources Association.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Lee, J. and Hasebe, Y. (2017). jWriter: Automated essay scoring system for learner using artificial intelligence mechanism - a trial using I-JAS -. In *2017 Autumn Meeting of the Society for Teaching Japanese as a Foreign Language*, pages 289–291, Niigata, Japan, November. The Society for Teaching Japanese as a Foreign Language.

Tanaka, M. and Kubota, S. (2016). Are standards for Japanese academic writing necessary?: Proposals based from analysis on essay organization. In *Proceedings of the 2016 Canadian Association for Japanese Language Education Annual Conference*, pages 263–274, Ontario, Canada, August. Canadian Association for Japanese Language Education.