IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol. 003, LIIR057
DOI: https://doi.org/10.52731/liir.v003.057

Implementation of Automated Feedback System for Japanese Essays in Intermediate Education

Huy Phan [*], Shinobu Hasegawa [*], Wen Gu [*]

Abstract

Traditional Automated Essay Scoring (AES) only provides students with a holistic score, unable to provide meaningful feedback on students writing. Holistic, structure, style, word, and readability are chosen from the 6+1 writing-trait theory to create an Automated Essay Feedback (AEF) for Japanese L1 students. By combining these rule-based traits with a data-driven model, we created a hybrid system that can automatically grade and give feedback to students. The system automatically identifies parts of student writing that need improvement, then recommends corrective and suggestive feedback. Our contributions are twofold: design a 5-writing-trait AEF for Japanese L1 students and implement the holistic corrective writing-trait.

Keywords: Automated Feedback System, Automated Essay Feedback, Question Answer System, 6+1 writing-trait, Japanese Language

1 Introduction

Corrective feedback (CF) with grades was shown to have positive effects on increasing student performance [1][2][3]. CF indicates where and how students can improve their writing, while grade provides an overall view of their performance. Grade and corrective feedback have a correlation with one another. The worse the student's grade, the more feedback is needed. But the relationship between the two is hard to justify because the semantic meaning is hidden deep in the feedback text, and it is difficult to compare the numeric score and the text. [4][5][6] built the Automated Scoring System (AES) only for grading the Japanese Language, but by using textual cosine-similarity [7] along with the students' scores, we can expand the AES to predict the scores and generate corrective feedback to create an Automated Essay Feedback System (AEF).

Even though AES is a good starting point to evaluate student performance, traditional Japanese AES [4], or English AES [8] have problems that their models use traits like total numbers or ratios, so the semantic meaning is lost, which results in low score prediction. Modern AES [5] improves score prediction by applying neural network models to create better semantic embeddings. But AES systems are limited to only providing students with overall scores, unable to show where and how the students can improve their writing. Furthermore, in Japanese AES [5] [6], the relationship between their systems and the writing theory is left untouched because the score from their system represents only a simple exist-a-certain-text-or-not trait, which is difficult to make a meaningful connection with any writing theory.

[*] Japan Advanced Institute of Science and Technology, Nomi, Japan

In 2003, the 6+1 writing trait [9] was used to teach US students from 3rd to 12th grade. Their goals were to remove the factory-like, uninspired essays and encourage students to put more effort into their writing. Research on the 6+1 model indicates a positive effect on the students critical thinking skills and writing [10]. Our long-term research goal is to create an AEF based on the 6+1 writing trait and discover how applying the writing theory can benefit the students. For the scope of this research, we design an AEF Open-Answer System and implement the first holistic trait from the 6+1 writing theory as our corrective feedback. Other traits like word, readability, style, and structure are categorized as suggestive feedback and will not affect the student's score. Two important tasks to create the feedback are score prediction and feedback generation. We will build these two models and evaluate the score prediction accuracy to [5] [6].

2 Automated Essay Feedback (AEF)

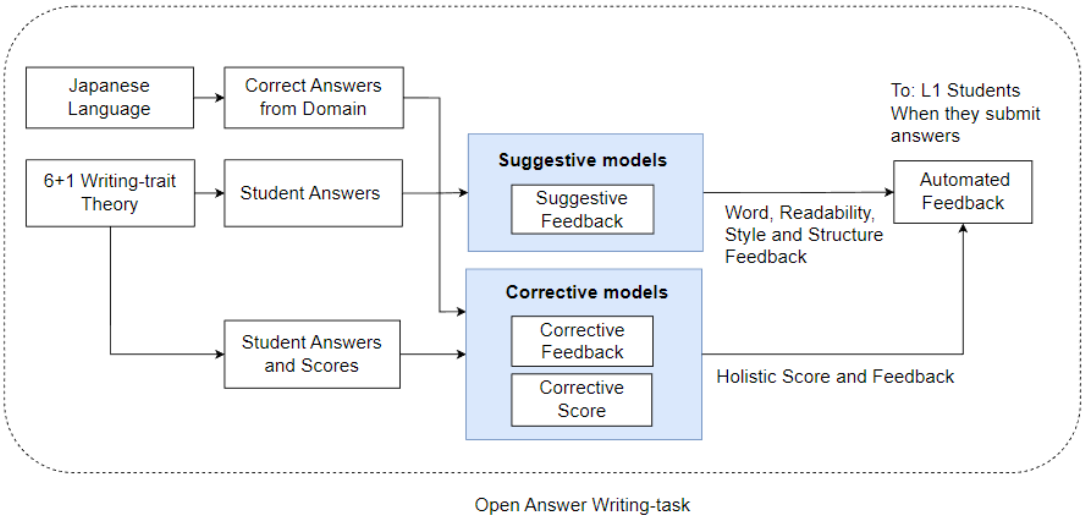

Figure 1: The structure of our AEF system

Technologies for Automated Feedback - Classification Framework (TAF-ClaF) [11] classifies 109 AFS into groups and abstracts them to a list of important characteristics. We then picked the characteristics suitable for our AEF Open Answer System and summarized them in Figure 1. Our system is a standalone technology, but the ideas can be transferred into any learning platform.

Japanese L1 students first choose a question from a list of predefined questions. Then they are required to write an essay as their answers. The required length of the answer is also predefined, and students should try to match that length. After finishing their writing, they proactively submit the answer into the system and receive automatically generated feedback. They will receive two types of feedback: suggestive and corrective feedback. Holistic scores and feedback are given by the corrective models. Word, readability, and style feedback are given to the students by the suggestive models. Only if the student's answer has more than two paragraphs, will structured feedback be given to them. These 5 traits are implemented using PyTorch. The backend is built with Python, and the frontend is built with HTML and JavaScript. The source code will be made publicly available at [12].

2.1 Open Answer Writing-task

Japanese students must participate in Japanese entrance exams when they go to junior or high school. And in their Japanese subject exams, 20% of the questions follow the Open Answer writing task format. They are required to write a short answer to the given question. After they submit, their answers are evaluated by comparing them to a correct answer extracted from a domain book. Our AEF system follows this same format to evaluate and give feedback to the students.

質問「こうした緊張したスタンスこそが饒舌な西洋文化を導いてきた」と
あるが、それはどういうことか。句読点とも七〇字以内で説明せよ。

Question: "This tense stance has led to the talkativeness of Western culture".
Please explain the sentence meaning in 70 words.

No	Model	Feedback
1	Peer Answer Score	模範解答「西洋人は基本的に他人は異人と見なす。」 Peer Answer: "Westerners have different perspectives."
2	Peer Answer Domain Answer Score	参考資料からの解答「西洋人は他人に自分とは異なる人間と見なすので、自分の考えに他人を同意させようと言葉を尽くして説得する。」 Peer / Domain Answer: "Westerners try to convince others to agree with them because they perceive themselves and others as different persons."
3	Peer Answer Teacher Feedback Score	先生からのフィードバック「西洋人はより良い解決法を見つけるため、他人との議論を激しく行う。そういう現象を解釈してください。」 Teacher Feedback: "You should explain that the Westerners argue to achieve a better solution for the both parties."

Figure 2: Open answer models

In an Open Answer System, answer, score, and teacher feedback can be used to create feedback for students. [5] [6] only use the answers in their models to predict the score. But our context is different because we also want to create feedback. So just using answers might be limited in what we can recommend to the students. In Figure 2, the first model uses peer answers as feedback. The best answer from the first model is the peer answer with the highest score. So, using the first model, students with high scores would not find any meaningful feedback for improvements. The second model utilizes the domain as the correct answer, meaning that the high-score student can still learn from the domain and improve. The first model relies on other peer answers to the same questions to be used as feedback. The second model relies on other peer answers and also the domain answers. The second model is useful in the case that a student already achieves a high score but needs a better reference to improve. And the third model is the best one, as it explicitly shows the students how to improve their writing with the teacher's feedback.

2.2 6+1 Writing-trait Theory

Define the prediction from the system as P, the student answer as A, the teacher feedback as F, the student answer's score as S, the weight of a given trait as W, the number of suggestive traits as h, and the number of corrective traits as $k - h$.

With the constraints $k \in [1,6], i + j > 0$.

Then we have the following formula which describes the 6+1 writing theory:

$$Feedback = \sum_{n=1}^{k} A_n = \sum_{n=1}^{k} F_n = \sum_{i=0}^{k-h} S_i \,.\, W_i \,.\, P_{Si} + \sum_{j=0}^{h} W_j \,.\, P_{Sj}$$

In layman's terms, the feedback for the students can come in two forms: the peer answers or the written feedback from the teacher. No matter the form, the feedback will be decided by the corrective and the suggestive models. The difference between the two models is corrective model comes with a score, while the suggestive model does not. If a system can give a measurable prediction for the student's grade, we call it a corrective feedback system. If the system, cannot generate the score, or their score is not measurable, we call it a suggestive feedback system.

Suggestive feedback is a set of rules extracted from the 6+1 writing-trait theory. Each of the rules comes with a weight to decide if it should be given to the student or now. And not all traits and trait characteristics are suitable for feedback, for example, convention and presentation traits. The convention trait in our context is assumed unnecessary because Japanese L1 students in intermediate education can already understand and use Japanese grammar well. The presentation trait can only be evaluated if the student is writing on a piece of paper. So, we reduce them to five crucial traits: organization, voice, word choice, and sentence fluency traits. Organization or structure means how well the student structures their long essays. Voice or style feedback means the unique style that the student applies in their writing. Word choice or word feedback, means how well the students use each individual word in their essays. Sentence fluency or readability means how well the students convey their ideas in sentences. Idea or syntactic is one of the traits of the 6+1 writing theory. It means how related the student's answer is to the question. The more the student's answer is similar to the question, the higher their score will be. The important characteristics for this trait already exist in the Riken Dataset [13] so we do not need to add any rule for this.

2.3 Quality of Feedback

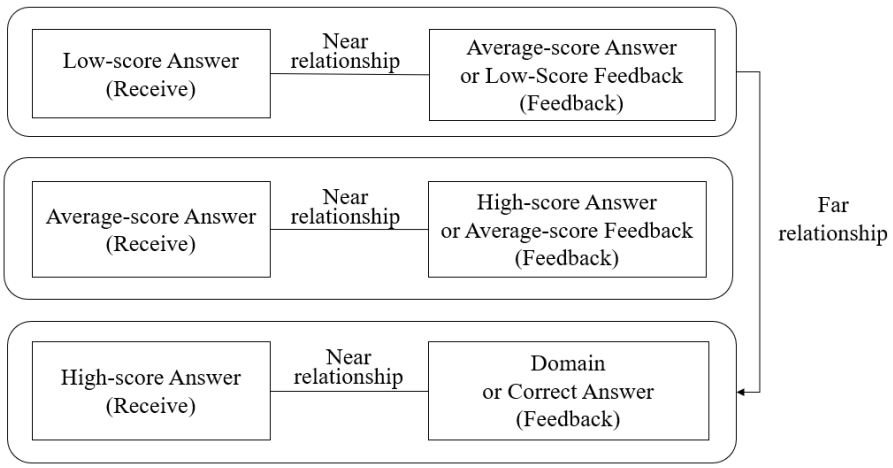

Figure 3: Answers and feedback's relationships

The traditional Question-Answer systems only provide students with the correct answer as feedback. But the learning curve between the student's answer and the correct answer might be too steep. The further the relationship between them, the more difficult for the student to learn from the feedback. By structuring the answers in the pairwise format, our system can be model-agnostic. Meaning that we can build the first, second, and third models in the same way as long as the dataset follows the pairwise format.

Our near relationship metric can be used to help ease the gap between the student's answer and the feedback. A near relationship is how close the semantic textual similarity (STS) and the scores between the receive and feedback group. The receiving group includes the students with generally lower writing scores than the feedback group. The feedback group can consist of the higher-score peer answers or the teacher feedback. This setup means that the students can receive the teacher's feedback or other peer answers as their references while making sure that their own answers can also be used as references for the other students as well.

The first metric shows the STS between the receive and feedback groups by using cosine-similarity [7].

$$NR_1 = Cos(x, y) = \frac{x.y}{||x|| * ||y||}$$

The second metric is calculated using the score between the receive group score and the feedback group score. This help identifies the elements in the feedback group.

$$NR_2 = A_{score} - P_{score} \leq \frac{1}{3}\max(A_{score})$$

For example, the high-score answer, and the correct answer might have a closer relationship than the low-score answer and the correct answer. By recommending the correct answer for the high score student, they might understand that relationship and improve their writing. But if we try to recommend the correct answer to the low-score student, the relationship might be too far, and they might not know where they are wrong. So, it is more beneficial to recommend the average-score answer to the low-score student as the relationship is closer.

2.4 Dataset

To build a Japanese AEF for the Open Answer task which can generate corrective feedback, the dataset needs to include at least two attributes: answer and score. The Riken Dataset [13] was created by conducting mock exams in a Japanese High School for 2 years. It has attributes like questions, answers, overall score, partial score, and annotated assessments. Total of seventeen questions, each with about 500-2000 answers. The answer length is short, around fifty words.

Riken Dataset attributes are enough to build the first model. But to build a second or third model, the dataset needs to also consist of the teacher feedback and the domain answer. The Riken dataset is good for predicting scores and generating feedback. For traits like structure, style, word, and readability, the dataset is not fit to create those types of feedback because it lacks the measurable scores of those traits. So, we cannot use those traits as corrective feedback with the Riken dataset.

But we can still use those traits as suggestive feedback, with no score indicators. In this research, we focus on building the first model. The second and third models will be for future research.

3 Corrective Models

Corrective models have two tasks: predicting the score of the student's writing (score prediction task) and measuring the near relationship using the cosine-similarity on students' answers, combing with the predicted score from the previous step to find a list of closely related answers to be given as feedback (feedback generation task).

$$Corrective = \sum_{n=1}^{k} A_n = \sum_{n=1}^{k} S_n \,.\, W_n \,.\, P_{Sn}$$

Using the Riken Dataset, we can only implement one trait - the holistic trait - from the 6+1 writing theory as our corrective model. Peer answers will be used for feedback. Weight is a constant and will be decided by conducting empirical experiments on the study groups. Then, the formula can be simplified as:

$$Corrective = Holistic = A_1 = S_1 . W_1 . P_{S1}$$

3.1 Score Prediction

3.1.1 Machine Learning baseline

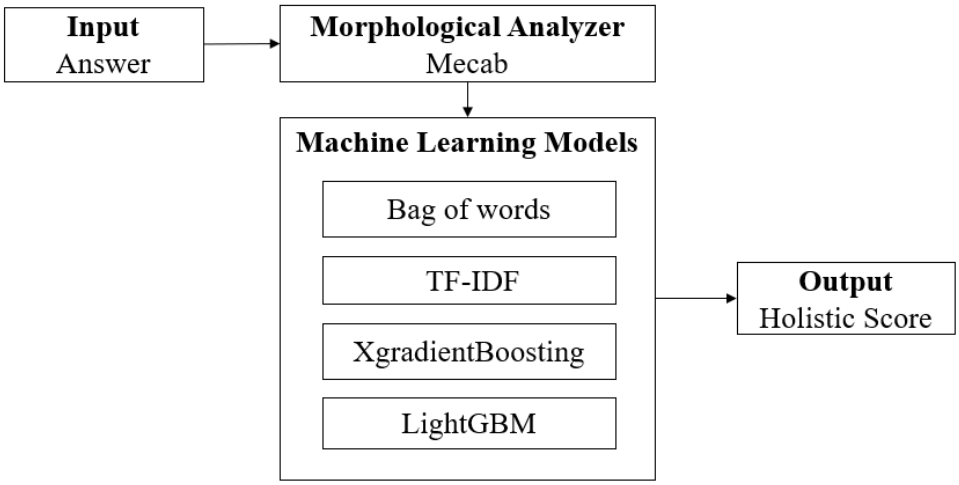

Figure 4: Machine learning models for score prediction

For the score prediction task, we experiment with a list of machine learning methods and pick the combinations that yield the highest accuracy among them. Our training features are limited to only the answer and the overall score. The Japanese Language is not separated with blank spaces like the English language, so the need of a morphological analyzer is required.

3.1.2 *Neural Network baseline*

Instead of using the Bag-of-words or TF-IDF model, the Bert model was used to extract the contextual meaning of the answer. And rather than building a linear layer on top of Bert for the regression score prediction task, we utilize the transfer learning characteristic of Bert and use other machine learning methods to handle those embeddings. Same as the machine learning baselines, the neural network model uses Mecab as the default morphological analyzer, and with the same number of training features.

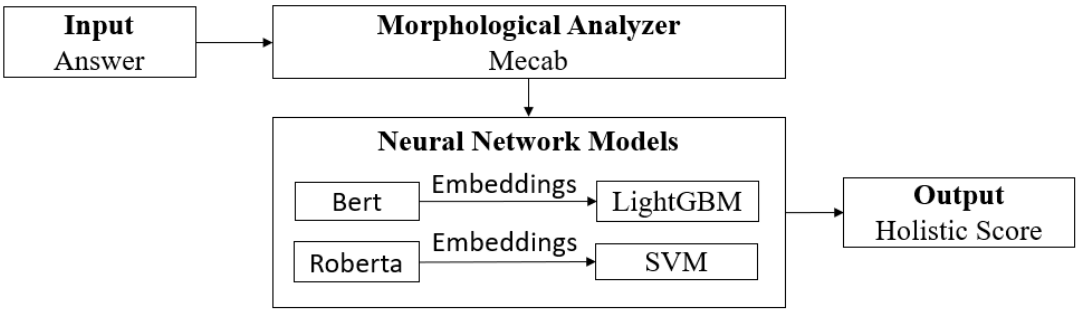

Figure 5: Neural network models for score prediction

3.2 Feedback Generation

The feedback generation task is done by using the predicted score from the previous step to identify the list of students' answers that is in the one-third upper range from the predicted score. Then use the Sentence-Bert model [7] to measure the STS on the embeddings generated from those students' answers with cosine-similarity. Those peer answers are now suitable to be used as feedback because they have similar semantic meanings to the input answer but achieved higher grading. If the students make changes according to the feedback, their scores will improve.

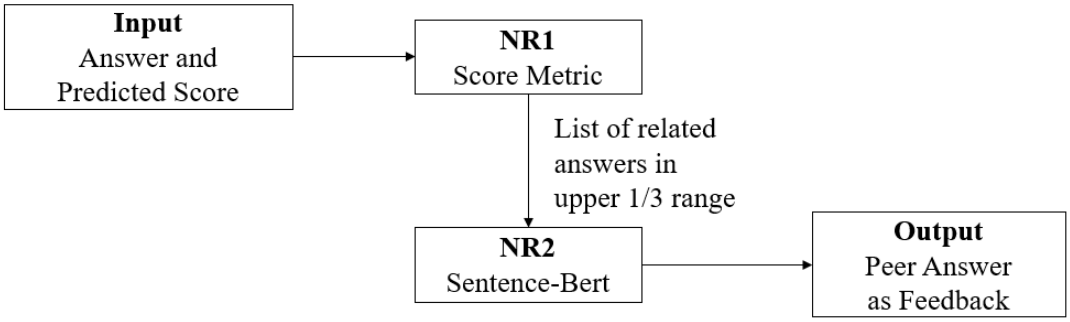

Figure 6: Feedback generation models

4 Suggestive Models

$$Suggestive = \sum_{n=1}^{k} W_n \,.\, P_{Sn} \;\; with\, k \in [1,6]$$

The actual number of suggestive models will be the remainder of 6 and the number of corrective models. The weight constants should be indicated by the empirical experiments to identify the importance of each rule. In our research, we use holistic traits as the corrective model, but in other systems or with other datasets, the holistic trait can be perceived as the suggestive model.

These following trait characteristics can be made into a list of predefined rules to be used as feedback. Or use them to find out the peer answers that satisfy these characteristics and use those peer answers as feedback.

Trait	Characteristics
Holistic (Idea / Semantic)	Explain, give directions, or extend idea. Clear main message with multiple detailed sentences.
Word	High-level, strong, expressive, and not repeated too many times.
Organization (Structure)	Focus on the topic. Follow a logical order. Use transition words. Have a beginning and an ending. Multiple sentences which show the development.
Readability (Sentence Fluency)	Write complete sentences. Variety of sentence lengths. Easy to read with expression. Begin sentences with different words.
Style	Write with a personal style. Use appropriate punctuation marks to enhance words. Write with a style like joyful, funny, fearful, angry, or serious.
Convention (Grammar)	Use capital, lowercase letters, periods, commas, exclamation points, question marks, and spell words correctly.

Table 1: Suggestive traits for writing

5 Preliminary Experiments

Score Prediction: Four methods were experimented for the baseline approach, it includes lightGBM, XgradientBoosting, Bag of Words, and TFIDF. After the experiments, we found that using lightGBM with Bag of Words result in the best prediction score among the four methods. Our neural network model is a work in progress, so after we implement it, we will compare score prediction results with our baseline model and [5] [6].

Model	Accuracy / R2 Score	RMSE
LightGBM + Bag of Words	**0.746**	**1.558**
XGradientBoosting + Bag of Words	0.740	1.576
LightGBM + TFIDF	0.725	1.618
XGradientBoosting + TFIDF	0.724	1.620

Table 2: Results of score prediction using machine learning methods

Model	Score Prediction Accuracy
Mizumoto 2019	0.87
Takano 2022	**0.88**
Our proposed Neural Network	0.65

Table 3: Results of score prediction using neural network methods

Our preliminary results compared to [5], [6] in the score prediction task are lower due to we are using only a few training features from the Riken Dataset. One important point is the use of the Justification Identification technique in the [5], [6] is not implemented in our research, which results in low overall score prediction accuracy.

Feedback Generation was applied to the Riken Dataset but was not able to be finetuned or measure the accuracy because Riken Dataset does not follow the pairwise format.

クエリ「西洋で生み出された理論や技法は、西洋的な個人や人間関係の在り方を前提にした人間観。」
Query: “Theories and methods developed in the West are based on the Western understanding of the human being, which is the Western way of individual and human relationships.”

Feedback	**Holistic Score**	**STS Score**
「神対人間、人間対自然、人間対人間という形で現される西洋文化の「対決」のスタンスのこと。」 “It refers to the Western culture’s ‘confrontational’ attitude, which manifested in forms of God vs man, man versus nature, and man versus man.”	4	0.0573
「西洋文化の基底には、自分の考えに相手を同意させる、神対人間、人間対自然、人間対人間という形で現れる「対決」があるということ。」 “Core of the Western society exists the ‘confrontational’ attitude, to persuade the other person to agree with their ideas. This "confrontational" attitude manifests as god versus man, man versus nature, and man versus Man.”	7	0.0767

Figure 7: Results of feedback generation

6 Conclusion

Three levels of feedback models can be created using the pairwise format. The three levels need to have distinct attributes like score, peer answer, domain answer, and teacher feedback to build a measurable AEF system that can recommend peer answer or teacher feedback based on close semantic meaning (STS) and close holistic meaning (score metric).

The generalized model of the 6+1 writing theory can be inherited not just for the Japanese Language nor the Intermediate L1 students but applied to other writing systems as well. By considering traits as corrective and suggestive, we can build a measurable writing system while still being able to provide students with other useful information about their writings. Another neat thing is when building a full writing system with suggestive models early, and then conducting some empirical experiments, we will know what traits we are looking for to build the corrective models.

By comparing the baseline results with [5], [6], we can foresee that the Justification Identification technique can have a positive effect on the score prediction accuracy, even though it is costly to annotate the dataset. The justification cues can be beneficial to transform suggestive models like the word trait model into the corrective model.

In future works, we will evaluate how the feedback from the 6+1 writing theory is useful for the improvement of student writing. And by conducting the empirical experiments, we will be able to create a list of baseline weights for each of the traits and their characteristics.

Acknowledgments

This study used the RIKEN Dataset for Short Answer Assessment, which was provided through the IDR Dataset Provision Service of the National Institute of Informatics.

References

[1] A. Hashemifardnia, E. Namaziandost, and M. Sepehri, “The effectiveness of giving grade, corrective feedback, and corrective feedback-plus-giving grade on grammatical accuracy,” International Journal of Research Studies in Language Learning, vol. 8, no. 1, Jan. 2019, doi: 10.5861/ijrsll.2019.3012.

[2] E. B. Page, “Teacher comments and student performance: A seventy-four classroom experiment in school motivation.,” J Educ Psychol, vol. 49, no. 4, pp. 173–181, Aug. 1958, doi: 10.1037/h0041940.

[3] C. van Beuningen, N. H. de Jong, and F. Kuiken, “The Effect of Direct and Indirect Corrective Feedback on L2 Learners’ Written Accuracy,” ITL - International Journal of Applied Linguistics, vol. 156, pp. 279–296, 2008, doi: 10.2143/ITL.156.0.2034439.

[4] T. Ishioka and M. Kameda, “Automated Japanese essay scoring system:jess,” in Proceedings. 15th International Workshop on Database and Expert Systems Applications, 2004., 2004, pp. 4–8. doi: 10.1109/DEXA.2004.1333440.

[5] T. Mizumoto 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, 2019, pp. 316–325. doi: 10.18653/v1/W19-4433.

[6] S. Takano and O. Ichikawa, "Automatic scoring of short answers using justification cues estimated by BERT," in Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), 2022, pp. 8–13. doi: 10.18653/v1/2022.bea-1.2.

[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.

[8] T. Zesch, M. Wojatzki, and D. Scholten-Akoun, "Task-Independent Features for Automated Essay Grading," in Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pp. 224–232. doi: 10.3115/v1/W15-0626.

[9] "Scholastic SC-0439280389-A1 Theory and Practice 6 Plus 1 Traits of Writing Guide, Grades 3 and Up." Scholastic Teaching Resources, Apr. 2018.

[10] A. A. Qoura and F. A. Zahran, "The Effect of the 6+1 Trait Writing Model on ESP University Students Critical Thinking and Writing Achievement," English Language Teaching, vol. 11, no. 9, p. 68, Aug. 2018, doi: 10.5539/elt.v11n9p68.

[11] G. Deeva, D. Bogdanova, E. Serral, M. Snoeck, and J. de Weerdt, "A review of automated feedback systems for learners: Classification framework, challenges and opportunities," Comput Educ, vol. 162, p. 104094, Mar. 2021, doi: 10.1016/j.compedu.2020.104094.

[12] "https://github.com/Hasegawa-lab-JAIST/huyphan-6-writing-trait-feedback."

[13] "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."