Master's Thesis

Implementation of Automated Feedback System for Japanese Essays in Intermediate Education

Phan Thanh Huy

Supervisor Shinobu Hasegawa

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

February, 2023

Abstract

Writing is a fundamental skill that plays an important role in student success in both academic and professional settings. Through writing, students can express their thoughts, ideas and knowledge and communicate effectively with others. This skill is essential for success in any field as it enables individuals to articulate their ideas clearly and persuasively. One of the most important reasons for writing is that it helps students develop critical thinking skills. Writing requires students to analyze, evaluate, and synthesize information, which helps develop analytical and problem-solving skills. Additionally, writing is critical to a student's academic success. At American school, students are required to write essays, and other written assignments as part of the curriculum. These assignments provide students with an opportunity to demonstrate their understanding of specific topics and develop their writing skills. It can help students organize their thoughts and present their ideas in a logical and coherent way. But Japanese exams are geared toward non open answer tasks as they are easier to score and fairer to grade. In these exams are homework, AES (Automated Essay Scoring) is a common method of scoring written answer using computer software. It is commonly used in education for grading student essays but can also be used in other areas. To analyze and evaluate written text, the AES system uses natural language processing and machine learning algorithms. It can be trained on a large dataset of previously graded essays to learn the characteristics of good essays and apply what it learns to grade new essays. AES can be used to provide student feedback, assist teacher. But AES can only provide students with a holistic score, unable to provide meaningful feedback on students writing. Automated Essay Feedback (AEF) can help provide those important feedback and be the important component of the learning process. Using those feedback, it can help students understand their performance, provides personalized and detailed information about their work, increases engagement and motivation, and aids in the development of critical thinking skills. But the feedback also needs to be chosen well before given to the students. Our research utilizes the 6+1 writing traits theory, which is widely used in American schools. The theory is widely accepted in teaching writing, assessing student writing, and providing feedback to students. The feedback from the theory is intended to be used as the comprehensive method for assessing student writing and can be used to provide feedback on the writing of students of all proficiency levels, from beginner to intermediate. Idea, structure, style, word, convention, and readability are chosen from the 6+1 writing-trait theory to create our AEF systems, that are suitable for Japanese L1 students. Idea trait is related to the content of the text and the quality of the ideas presented.

This includes the relevance, originality, and development of ideas within the essay. Structure trait is related to sentence structure and coherence. This includes the logical flow of ideas, the use of transitions, and the overall coherence of writing. Style trait refers to the students' voice and tone when writing. Word trait is related to the students' use of vocabulary and language. This includes the accuracy, meaning and appropriateness of the words used in the writing. Convention trait refers to the students' adherence to the rules of grammar, spelling, and punctuation. It includes the students' ability to use capitalization, punctuation, and grammar correctly and consistently. And the final Readability trait refers to the general ease of reading and understanding what is written. This includes the ability of how students can communicate their ideas clearly and concisely, and the use of formatting and visuals to enhance readability. By combining these traits with a data-driven model, we created a system that can automatically grade and give feedback to students. The system automatically identifies parts of student writing that need improvement, then recommends feedback to the student. The feedback can come into two form, corrective and suggestive feedback. Suggestive feedback is a type of feedback that gives students advice and suggestions on how to improve their writing. It is different from corrective feedback, which simply tells the students whether their writing is good or bad. While suggestive feedback is preferable, it is not always easy to give or receive. It takes time, effort, and practice to provide feedback that is both specific and actionable while remaining non-obtrusive. Corrective feedback has both advantages and disadvantages. The advantage is that it provides a clear picture of a student's overall achievement and is easy to understand for students, parents, and teachers. The downside is that it doesn't provide specific information about what students need to do. It also does not give students the opportunity to take control of their own learning process. While it can provide a judgment or assessment of the quality of a student's writing and give students a general idea of how well they are doing in their writing, corrective feedback is best to be used in conjunction with other types of feedback such as suggestive feedback, to give students specific information on what they need to improve. Our contributions in this research are twofold: design a 6+1 writing-trait AEF for Japanese L1 students and evaluate a new feedback type using peer answer as the feedback. We propose that every AEF system should partially relies on AES system, so the feedback can be measured based on the student's score metric. The result from our system shows that using peer answer as feedback can lead to the improvement of student writing and students prefer human-like feedback more than peer-answer feedback as the latter often lack the context and explanation.

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

Contents

This dissertation was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and University of Information Technology, Vietnam National University of Ho Chi Minh City.

List of Figures

List of Tables

Acknowledgements

This research is Master Thesis for JAIST, from the 1-1 Program with University of Information and Technology Vietnam. We would also like to express our sincere gratitude to Prof. Hasegawa for teaching us valuable knowledge in the field of education. Without your guidance, this Master Thesis would not be finished. We have made great efforts to implement this topic, but due to limited knowledge, experience and time, this project still has many limitations and shortcomings. We are looking forward to receiving many comments and suggestions to improve the content of this project.

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.

Chapter 1
Introduction

1.1 Background

Writing is an important aspect of language that performs several important functions in communication and expression. It allows us to communicate our thoughts, ideas, and feelings to others in a clear and structured way, and it helps us retain and transmit knowledge and information across time. space and space. One of the main functions of writing is to facilitate communication. Whether we write letters, emails, or text messages, we use writing to communicate our thoughts and ideas to others. Writing allows us to share information and communicate with people who may not be physically present, and it allows us to communicate with people who speak different languages or in different parts of the world. Another important function of writing is to record and store information. Written language serves as a medium for recording and storing knowledge, whether in the form of a scientific article, a historical document, or a literary work. Writing allows us to preserve information over time and pass it on from generation to generation, ensuring that knowledge and experiences of the past are not lost. In addition to its practical functions, writing also has an important role in self-expression and creativity. Through writing, we can express our thoughts, feelings, and opinions in a personal and authentic way. Whether we are writing a poem, a story or a personal essay, we can use writing as a vehicle for self-expression and artistic creation. Another important aspect of writing is its ability to help us organize and structure our thoughts. When we write, we are forced to put our ideas into words and put them together in a logical and coherent way. The process of organizing and structuring our thoughts can be beneficial in many ways. First, it helps us clarify our own thoughts and ideas. When we have to write down our thoughts, we are often forced to think more critically and deeply about what we are trying to convey. It can help us better understand our own ideas and communicate them more effectively to others. Second, writing can be a powerful tool for learning and critical thinking. As we write, we are asked to seek and gather information, evaluate and analyze sources, and form and defend our own arguments. This process can help us develop critical thinking skills, better understand and interact with the world around us. Writing is also an important tool for social and cultural expression. Throughout history, writing has been used to record and transmit cultural traditions, beliefs, and values.

It has played an important role in preserving and transmitting the collective knowledge and experience of a society. In recent times, writing has been used to record and preserve the stories, traditions, and views of marginalized and under-represented groups, such as indigenous and underrepresented communities. minority culture. In addition to preserving cultural traditions, writing can also be used to challenge and change social and cultural norms. Throughout history, writers have used their words to advocate for social justice and raise awareness about important issues. Whether through poetry, fiction or non-fiction, writing has been a powerful tool for activism and social change. In today's digital age, writing continues to play an important role in cultural and social expression. The Internet has provided a platform for people from all over the world to share their thoughts, experiences and views with a wider audience. Writing can be used to connect people from different cultures and backgrounds, while promoting understanding and dialogue. In summary, the role of writing in language is multifaceted and has decisive significance for human communication and understanding. It allows us to share information and knowledge, express ourselves creatively, and interact with the world around us. The written word has also played an important role in the preservation and transmission of cultural traditions and values as well as in the defense of social justice. In the digital age, it continues to be a powerful tool for connecting people of diverse cultures and fostering understanding and dialogue.

Corrective feedback (CF) refers to the feedback that teachers or peers give to learners on their written work, such as essays, assignments, or exams. CF is an important tool in language learning and teaching. The feedback helps learners identify and correct mistakes in their writing and improve their overall language skills. There are several reasons why CF is important in language learning. First, CF helps learners identify and correct mistakes in their writing. This is important because errors can interfere with the reader's understanding of the writing. By pointing out errors and suggesting corrections, CF helps learners improve the accuracy and clarity of their writing. Second, CF helps learners develop their language skills. By providing feedback on language usage, teachers help learners understand the conventions of the language they are learning. For example, CF helps learners understand the correct usage of tenses, word order, and other language conventions. Third, CF helps learners develop critical thinking skills. By analyzing and correcting mistakes in writing, learners can improve their critical thinking and problem-solving skills. CF alone is an important tool in language learning and teaching. It helps learners identify and correct mistakes, improve their language skills, and develop critical thinking skills. Nevertheless, CF along 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.

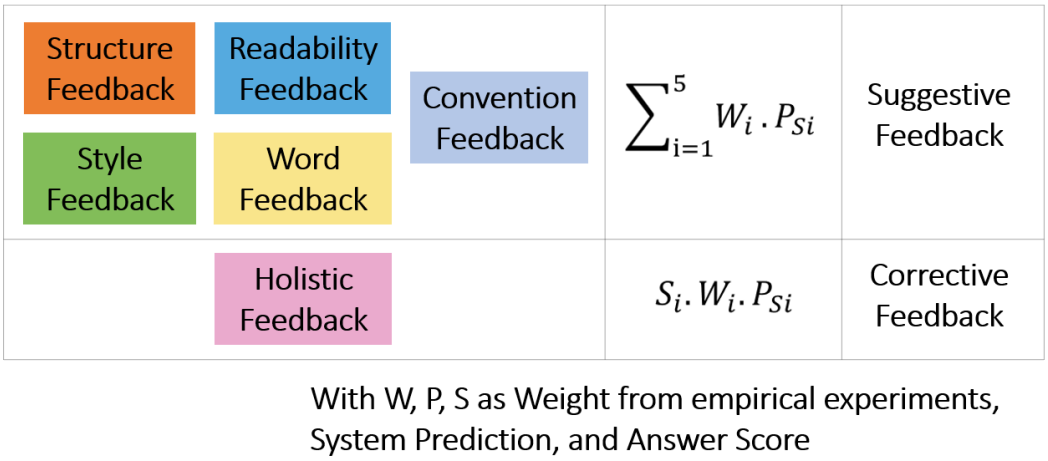

Figure 1: Six traits feedback for Japanese high school students

In 2003, the 6+1 writing trait [9] was used to teach US students from 3rd to 12th grade. Figure 1 shows the six important types of feedback in the 6+1 writing theory. In our research, we group the structure, readability, convention, style, and word feedback into suggestive feedback group. The corrective feedback group only contains the holistic feedback. The goals of the six feedbacks are 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].

1.2 Problem Statement

Japanese high school students are among the most academically successful and hardworking students in the world. They are known for their dedication to their studies and strong work ethic, which is reflected in their high-test scores and impressive college enrollment rates. One of the reasons for this academic success is the importance of education in Japanese culture. From an early age, students are taught the values of hard work and perseverance, and these values are reinforced throughout their school career. In high school, students often stay up late doing homework and preparing for exams, requiring long hours of study. Japanese high school students are known for their dedication to their studies and strong work ethic. Supported by a culture that values education, they a achieve academic and personal achievements even when faced with challenges.

To encourage Japanese high school students to use greater creativity in their writing assignments while also providing them with a score that indicates whether their answers are correct, we design our system based on the writing tasks that the students may encounter in their exams. In the system, the 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 structure feedback be given to them.

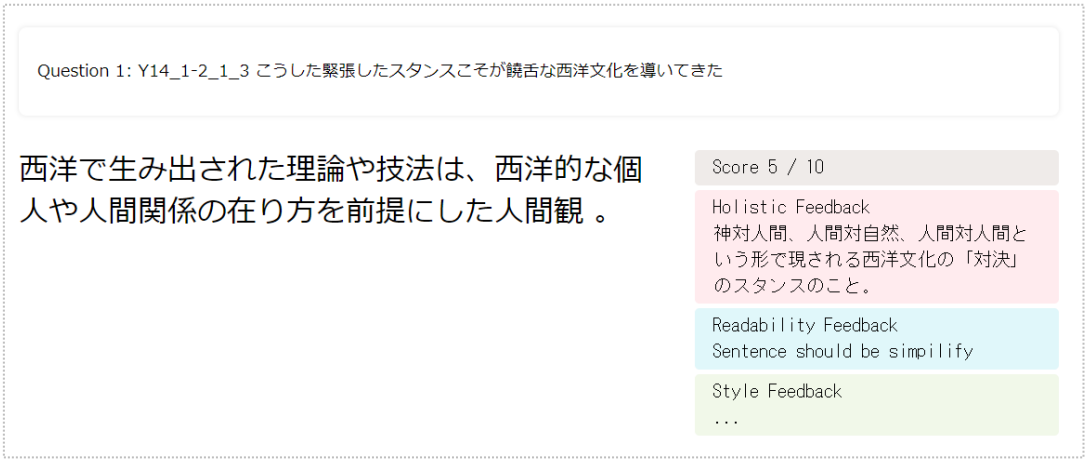

Figure 2: User interface of our AEF system

Figure 2 is the user interface that we develop to help to student write and get the feedback from our system. We highlight the score and different feedback in distinct colors to make them more noticeable by the students.

Our research aims at developing an automated feedback system for Japanese students in intermediate education to promote the use of writing exercises in Japanese education, improve the student's writing through the use of feedback, cut down the grading and feedback workloads of teacher. To achieve this goal, I began with the following research questions:

RQ1. What feedback to include in Japanese automated essay feedback system?

RQ2. What feedback improve writing of Japanese students in intermediate education?

RQ3. What method can improve score prediction accuracy of an essay?

1.3 Research challenge

Japanese intermediate education lacks the writing exercises and writing exams. Even if more writing exercises and exams are introduced, teachers would still be overwhelmed with huge numbers of exercises to grade and give feedback on. And with no criteria on how to grade and feedback, teacher cannot always give a fair result to the student. After setting up the feedback criteria, the need of finding which feedback results in improvement of student's writing score is also very important. At the time of this thesis, automated feedback system for open-answer task in Japanese or English language are still not well researched.

1.4 Thesis structure

The thesis comprises 6 chapters. In the first chapter, we present the background of current learning approach, the corresponding challenges and then condense the research problem into the research questions to be addressed. Next, in chapter 2, we present the literature and technologies to which we refer in order to seek for answers to the research questions. Then, in chapter 3, we propose a formalized model for 6+1 writing theory that can consider the quality of feedback to make it suitable for Japanese students. Then we will discuss how we conduct the system and educational experiments in the chapter 4. In the chapter 5, we will see the preliminary results from the models in the previous chapter. Finally, chapter 6 provides the research conclusions and future research ideas.

Chapter 2

Related Works

2.1 Hirao 2020 model

The objective of Hirao model is to improve the writing prediction accuracy in AES system. By introducing organization, content, language, and holistic traits, they are able to identify the important aspects of a student's writing. In our research, even though we use different training features than Hirao model's, the traits in their works have many similar aspects to the 6+1 writing theory, which can support the importance of 6 traits feedback in AEF system.

Figure 3 shows the overall of Hirao model with the use of their training features to prediction the score for a student answer. Their model, with the use of Bert, focus on counting the total number of characters, morphemes, commas, sentences, and paragraphs, as well as examining the number of morphemes in common with the essay prompt. The organization of the essay is given the most weight, with the number of paragraphs being the most important factor. The language in the essay is also analyzed, including the ratio of hiragana, katakana, and kanji.

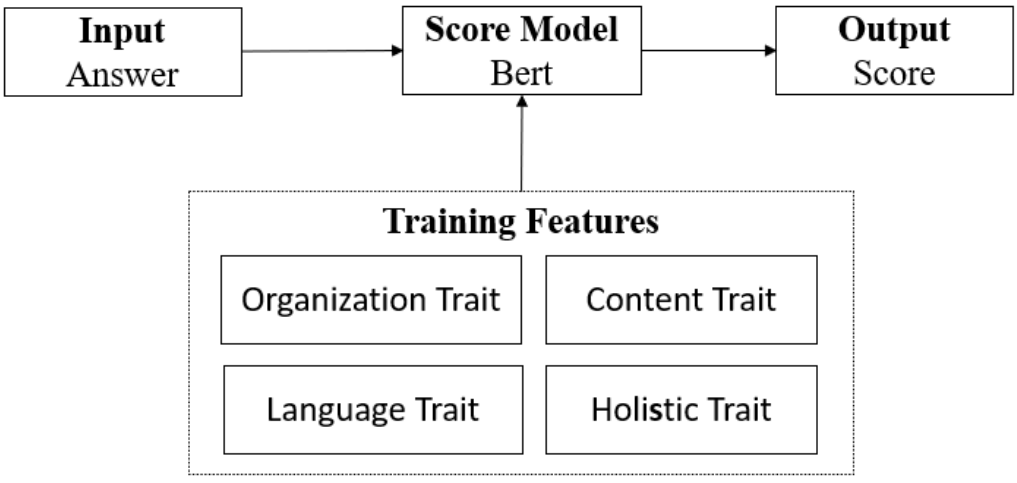

Figure 3: Hirao's score model

2.2 Mizumoto 2019 model

Mizumoto [5] has the most influence on the score prediction task of our model. The dataset that they are using are also used in our research, but in a slightly different ways

as in the Mizumoto model they often focus on getting the highest score prediction accuracy possible in an AES system. The main difference is that our AEF system focuses on generating the feedback based on the score and other attributes as well. Their architecture also uses LSTM model while our model uses transformer model like Bert for the score prediction task.

Mizumoto main contribution is to use Justification Identification features as an analytic criteria. In figure 4, we can see how Mizumoto combine the Justification Identification features with other known features like partial score, overall score, and student answer, for evaluating the student essay. The analytic score includes three smaller criteria (A, B, and C) that will give student a score if they contain certain keyword, as well as an additional criterion (D) that will result in a deduction of points if there are any misspellings or minor flaws. The dataset used in their research are scored manually by expert annotators which can result in high accuracy.

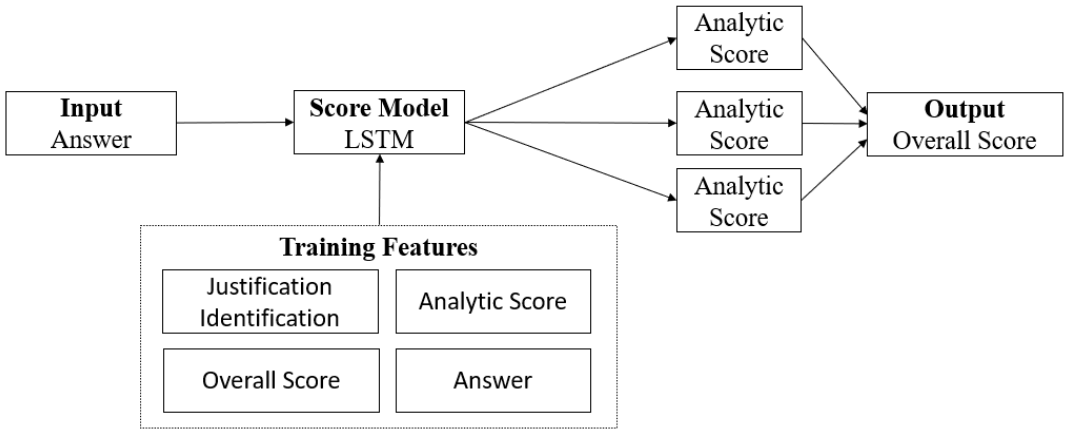

Figure 4: Mizumoto's score model

2.3 Technologies for Automated Feedback (TAF-ClaF)

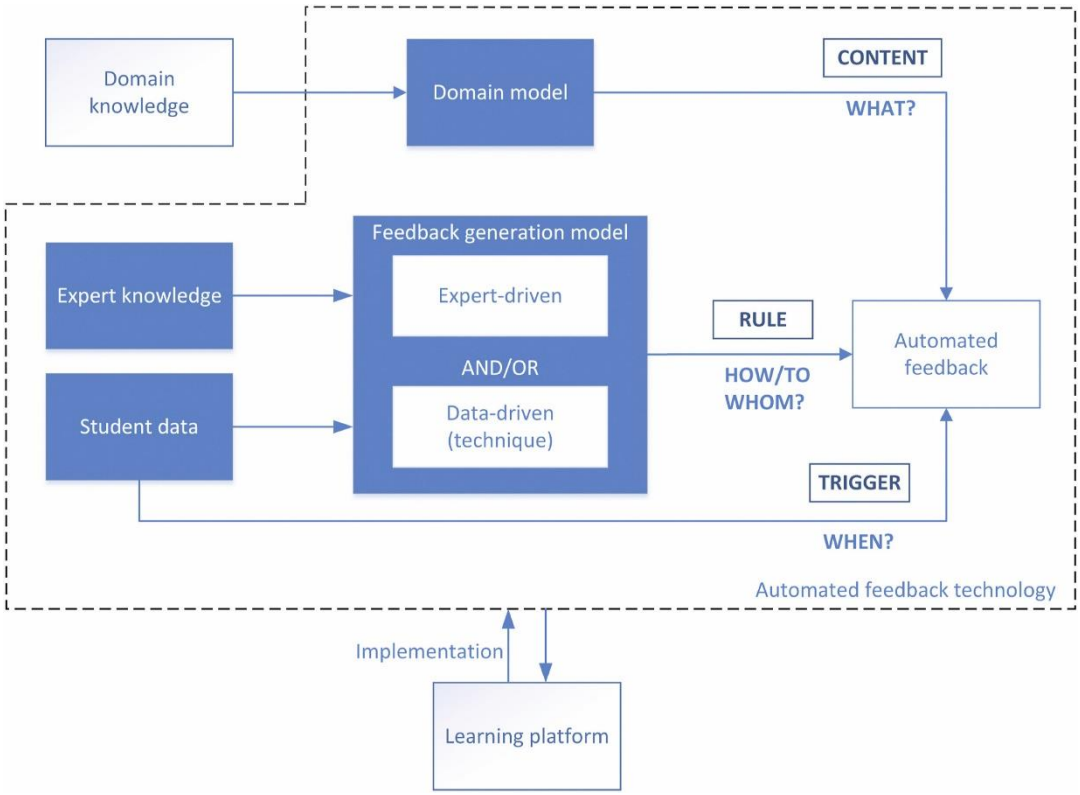

Figure 5: TAF-ClaF System

Technologies for Automated Feedback - Classification Framework (TAF-ClaF) [9] was used as a guideline when we build our AEF system. In figure 5, the domain knowledge is the content that an AEF system ultimately suggests to the student. Our proposed model approaches this a little bit different as we will have 3 different ways of giving the feedback to a student, not just from the domain. We will discuss more about this difference in the latter part. Another thing in the TAF-ClaF that we uses are the feedback generation model, they separate the models into 2 main types, expert-driven and data-driven model. In this research, we use the expert-driven as the rule-based suggestive model and data-driven as the corrective model.

TAF-ClaF contributions are by classifying 109 Automated feedback system into groups and abstracts them to a list of important characteristics. The classification framework consists of four main components: architecture, feedback, educational context, and evaluation, each characterized by several dimensions. TAF-ClaF was developed in two iterations using the design science research approach introduced by Hevner. The relevance cycle comprises the initialization of research, i.e., the problem that needs to be addressed and the means of solution evaluation. Next, the rigor cycle embodies the existing methods or frameworks helpful to construct and evaluate the solution. Finally,

the design cycle is the key component of design science research, representing the process of solution development and its evaluation. Ultimately, we use TAF-ClaF to answer the most important question: “What feedback to give to the student?”. By combining TAF-ClaF with the 6+1 writing theory, we are able to shed some light on how to combine a frequently used writing theory in the US and mold it to be more suitable for Japanese intermediate education.

2.4 BERT

BERT (Bidirectional Encoder Representation from Transformers) [11] is a language model developed by Google in 2018. It is a transformer-based architecture that handles input words in two directions, meaning is that it considers the context of words before and after the current word, rather than the context. of words that precede the present word as in traditional language models. One of BERT's key innovations is the ability to understand the meaning of a word in context, known as contextual embedding. This is done through the use of attention mechanisms, which allow the model to focus on specific parts of the input when making predictions. BERT is trained on a large data set of unannotated text, called a “Hidden Language Model” (MLM), where some words in the input are randomly hidden and the model predicts hidden words based on surrounding context. BERT has achieved industry-leading results on several natural languages processing tasks, including language translation, question answering, and text classification. BERT was initially trained on English text but can also be customized for other languages, including Japanese. Fine-tuning BERT for Japanese requires additional preprocessing steps, such as segmenting Japanese text into individual characters or words. For the score prediction task in our study, we employ a pre-trained Japanese Bert [12] (cl-tohoku/bert-japanese) from Tohoku University.

Roberta [13] is a variant of BERT developed by Facebook in 2019. It is based on the same transformer-based architecture as BERT but trained on a larger data set and use a different training process. Roberta is designed to improve the original BERT model in several ways. First, it was trained on a dataset three times larger than the one used for BERT, allowing it to learn more about the structure and patterns of the language. Second, it uses a different training process called "dynamic masking", which randomly masks multiple words in each input sample instead of just one, forcing the model to rely more on context and less on specific words. The Japanese pre-trained Roberta (cl-tohoku/roberta-base-japanese) [14] in our research also comes from Tohoku University.

The reason we choose BERT, and Roberta in our research is because they have achieved state-of-the-art results on several natural language processing tasks and have

been used in a variety of applications. They are especially effective in tasks that require a deeper understanding of languages. The two Japanese pre-trained models of Bert and Roberta can help improve the score prediction accuracy versus the LSTM model of previous work.

2.5 Semantic Textual Similarity

Semantic Textual Similarity (STS) [15] is a measure of the degree to which two pieces of text convey the same meaning. It is an important concept in natural language processing and has many applications, including information retrieval, machine translation, and automatic es-say grading. One of the main challenges in determining the semantic similarity of text is that the meanings of words and phrases can change depending on their context. Therefore, determining the semantic similarity of two pieces of text requires the ability to understand the context in which words and phrases are used and determine their underlying meaning. There are several approaches to measuring the semantic similarity of text, including the use of word embeddings, which are mathematical representations of words that capture their meanings and their relationships. with other words. Word embedding can be used to compare the meanings of words and phrases in different pieces of text and to determine how similar they are. Another approach to measuring the semantic similarity of text is to use semantic analyzers, which are software tools that analyze the grammatical structure of a text and determine its underlying meaning. STS can be used to extract meaning from text in many languages and can be particularly useful for comparing the meaning of text in different languages.

Sentence BERT (SBERT) [15] is a natural language processing model developed to better understand contextual relationships between words in a sentence. It is a variant of the BERT model. SBERT is specifically designed to encode contextual relationships between words in a sentence, rather than the individual words themselves. This is important because the meaning of a word can change depending on the context in which the word is used. The context provided by other words in the sentence helps clarify the meaning of the word. To do this, SBERT uses a transformer architecture that considers the order of words in a sentence. It has two main components: an encoder and a decoder. The encoder takes the input sentence and converts it to a fixed length vector representation, which is called embedding. The decoder then takes this integration and makes a prediction for the current task. One of the main advantages of SBERT is that it can be tailored to many NLP tasks. This means it can be trained to perform a specific task, such as sentiment analysis or language translation, by weighting the model with a small amount of labeled data. This allows SBERT to achieve high performance on these tasks

without the need for large amounts of labeled data, which can be expensive and time-consuming to obtain.

In our research, STS metric calculated from SBERT of all the student answer can help us group the related answers together. Then combine with the predicted student score, the system can generate a feedback based on related peer answers.

2.6 GPT-3

GPT-3 [16] is a neural network-based language processing model that generates human-like text through unsupervised learning. It has been trained on a massive amount of diverse text data and can generate a variety of text, including natural language, code, and poetry. It can be tuned for a variety of natural language processing tasks, including language translation, question answering, and text summarization. In our research, we want to use this model to generate the human-like (suggestive feedback) and do comparison with the effectiveness of peer answer (corrective feedback). SBERT in the previous section can help us generate the peer answer feedback, and the ChatGPT API [17], which utilizes the GPT-3 model, can help us generate the human-like feedback in our research.

Chapter 3

Proposed Model

3.1 Approach

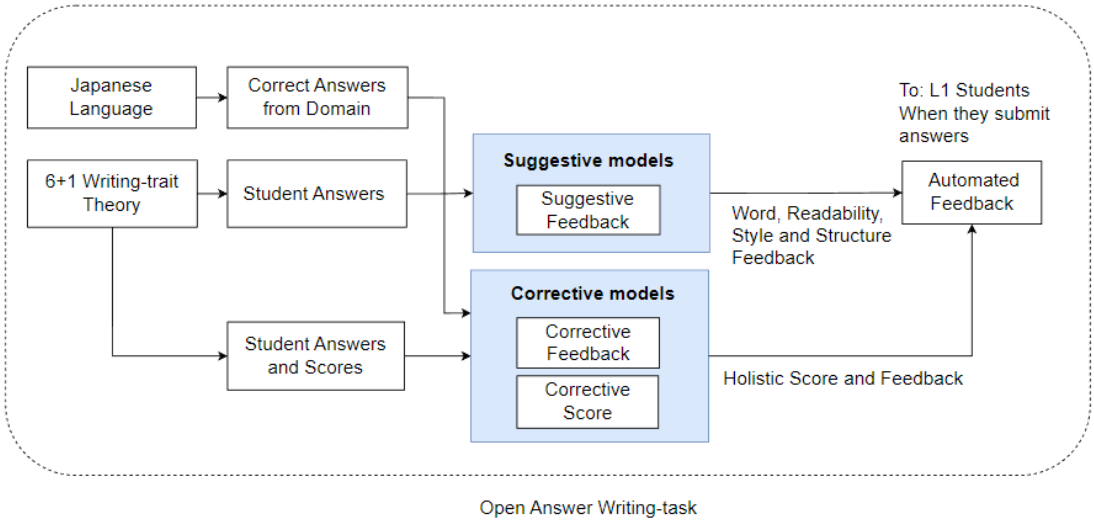

Figure 6: Proposed AEF system

In our research, we picked the characteristics from TAF-ClaF to build our AEF Open Answer System and summarized them in Figure 6. The two most important parts from TAF-ClaF framework are: domain, and feedback models. In their framework, the feedback came from a domain knowledge, but in our proposed models we make some adjustments by introducing feedback than can both come from domain, score, and peer answer. They are the inputs of our feedback model. We further break the feedback generation model into suggestive and corrective model. These are similar to the original design of TAF-ClaF where they introduce expert-driven and data-driven model. Then in the output of the AEF system, it also matches the original design of TAF-ClaF framework, where the students can request new feedback for their answers whenever they submit it.

Our system is a standalone technology, but the ideas can be transferred into any learning platform. These models 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 [18].

3.2 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 [19] 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. All grading data is stored in JSON format. Some questions are graded independently by two graders, and the annotation results are saved in a separate file. Figure 7 shows the most appear words in the dataset. There's indication that these appear words match the questions and criteria in the dataset.

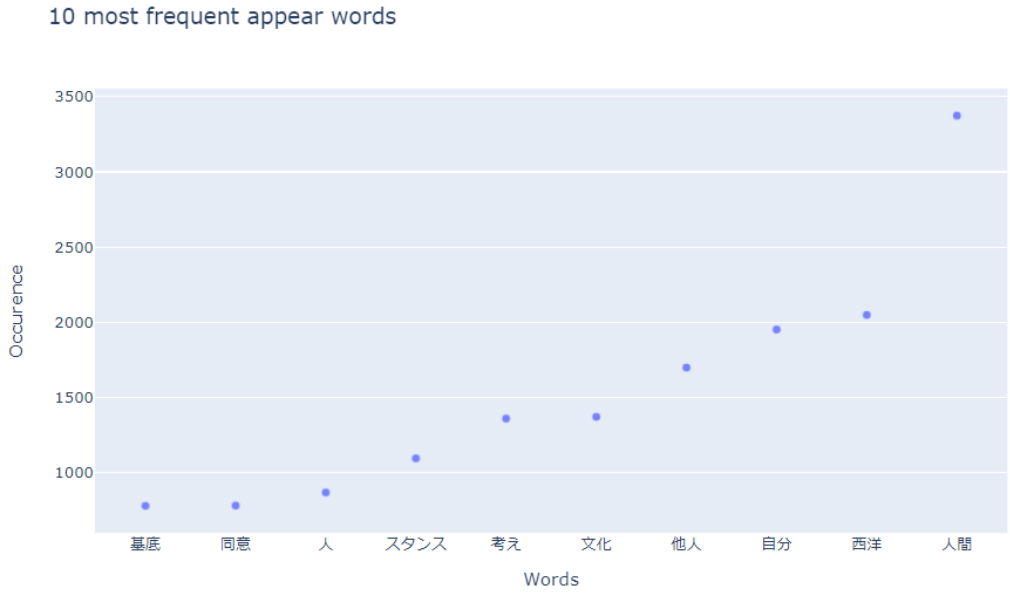

Figure 7: 10 most appear words in Riken dataset

Specifically, the data record has features A, B, C, D, E as evaluation criteria annotations. The i-th annotation (0 or 1) corresponds to the i-th morpheme of "mecab". Each letter (A, B, C, D, E) is associated with a rating element defined in the rating rubric. Response text di-vided into character segments C_A, C_B, C_C, C_D, C_E The characteristics are annotations of the evaluation criteria. The i-th annotation (0 or 1) matches the i-th character of 'Char'. Each letter following "C_" (A, B, C, D, E) is associated with an evaluation item defined in the evaluation criteria. A feature ID is a serial number in a record. The score function is the total score. Basically, it is calculated by adding the points of the following items. A total score is always calculated by adding item scores as

described below. How-ever, if the rating is negative, it is set to 0. A_Score, B_Score, C_Score, D_Score, E_Score, Miss_Score, EOS_Score are scores. These correspond to the items that are added in the scoring criteria, and Miss_Score and EOS_Score are the scores that correspond to the items that are subtracted has been manually corrected. Figure 8 is an example of a record, with the student overall answer score is 8, and the partial score of A_Score, B_Score, and C_score are in turn 0,0,8.

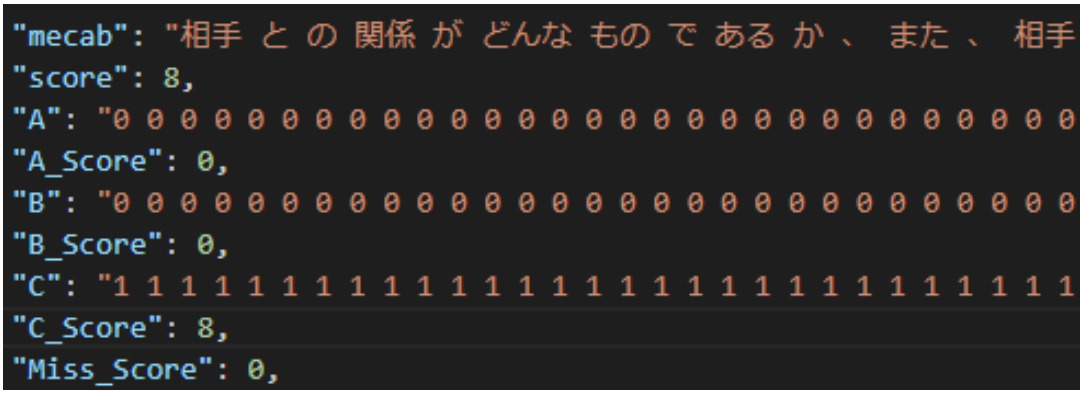

Figure 8: Riken dataset's attributes

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.

3.3 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 9: Three types of feedback

Figure 8 is a list of all possible holistic feedback model that can be created using our method. The similarities are that they all follow text pair formats. In the Open Answer System, the attributes from these models like peer answer, domain 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 above figure, the first model uses peer answers as feedback. But 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.

3.4 Quality of Feedback

Feedback is an important part of the writing process, especially for students. It helps students identify strengths and weaknesses in their work and make necessary corrections to improve their writing. Feedback can help students understand their work better. When students receive feedback on their writing, they can see their work from a different perspective. This will help them identify areas where they need to refine ideas or make changes to improve the overall consistency and effectiveness of their writing. By making

the necessary changes based on the feedback that they received, students can improve their writing skills and become more proficient over time. By learning to identify and correct mistakes and weaknesses in their writing, students can improve the overall quality of their work. Feedback also helps students write with more confidence. When students receive positive feedback on their writing, it boosts their confidence and helps them feel comfortable and competent. On the other hand, receiving constructive feedback also helps students develop a growth mindset and see writing as a skill that can be improved with practice and effort.

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.

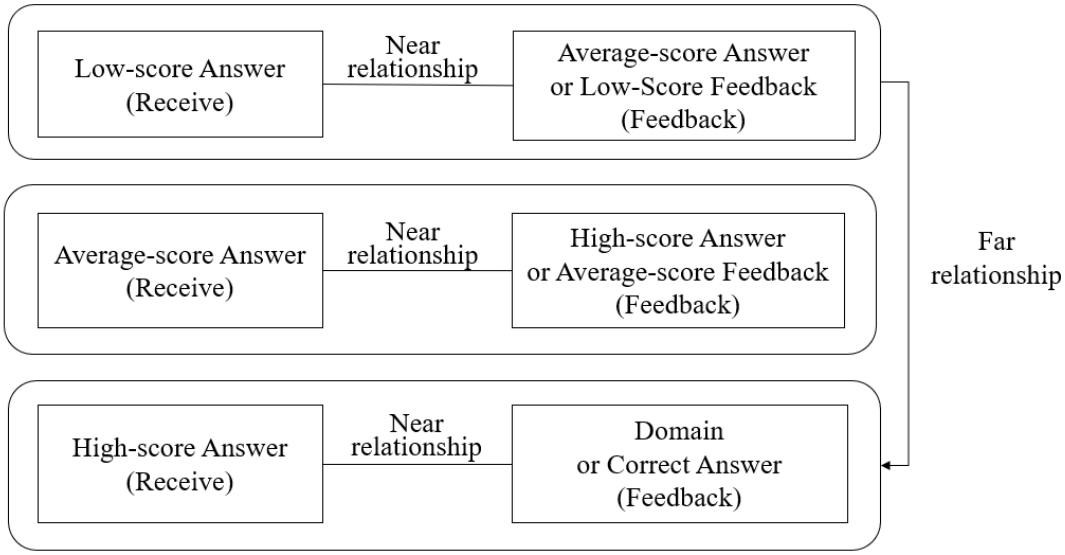

Figure 10: Answer and feedback relationship

Figure 10 shows our representation of near and far relationship. The 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.

3.5 An Ideal Design for Suggestive Models

An ideal design of an automated feedback system should include all 6 traits from the writing theory. But in the scope of this research, we will only focus on building the corrective feedback model. Nevertheless, the characteristic of the suggestive model is also described even though not actually being implemented to help us have a look of what characteristics that an automate feedback system should satisfy.

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

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. Figure 13 shows our ideal system with the use of 6 type of feedback.

Table 1: Characteristics of 6+1 traits

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.

Holistic or idea trait is the basis of all writing. Without a clear, well-developed idea, the essay can lack focus and direction, making it difficult for readers to understand and interact with the content. The importance of ideas in the writing can be seen in many ways. First, a strong idea drives the essay, providing a clear focus and direction for their writing. It gives students a reason to write and helps guide the structure and organization of content. In addition, a well-developed idea can help attract and engage the reader. When the student has a clear and compelling idea, the reader can more easily connect with the content and feel invested in the essay. On the other hand, writing without a clear idea can be difficult to capture the reader's attention, as it may lack the depth and content needed to sustain interest. Ideas are also important in writing because they can help convey an important message or theme. Through developing and exploring an idea, a student can communicate important ideas and concepts to their audience, whether through argument, storytelling, or other means. A unique, well-thought-out idea can help a student’s work stand out and be more memorable.

Word trait is an important aspect of writing for many reasons. The words students choose can have a significant impact on the tone and writing style. Different words can convey different emotions, meanings, and degrees of formality. Choosing the right words can help students effectively convey the tone and message they want to their audience. The words can affect the clarity and effectiveness of student’s writing. Using precise, specific words can help students to communicate ideas clearly and make their writing more engaging and persuasive. On the other hand, using vague or overly complicated words can make their writing difficult to understand and may turn off readers. The word

choice also affects the readability and flow of the essay. Using a variety of words and sentence structures can make their writing more interesting and engaging, while using the same words and structures over and over can make the writing monotonous and boring. boring. Choosing the right words can help create a smooth and fluent reading experience for the audience.

Organization or structure trait is an important part of writing that helps organize and present ideas clearly and logically. It helps readers understand and follow the flow of thought, and helps students communicate their ideas effectively. There are many different types of structure that can be used in writing, including temporal structure, spatial structure, and thematic structure. The temporal structure organizes information in a linear fashion, with events occurring in the order in which they occur. This is often used in narratives or writing about historical events. Spatial structure arranges information by location, with ideas presented according to their physical relationship to each other. This is useful when describing a parameter or when providing instructions. Thematic structure organizes ideas around a central theme or idea, with each paragraph or section exploring a different aspect of the topic. Regardless of the type of structure used, it is important to have a clear and logical flow of ideas. This can be achieved using transition words and phrases, which help connect one idea to the next and guide the reader through the writing. It is also important to consider the reader when deciding on the appropriate structure for a writing. Different audiences may have different needs and expectations, and the structure should be tailored to meet those needs. In addition to helping to organize and present ideas, structure also helps to improve the overall clarity and consistency of the writing. When paragraphs are presented in a clear and logical manner, it is easier for the reader to understand and follow the main points made. This is especially important when writing academic or technical writing where complex ideas can be presented.

Readability trait is an essential aspect of writing that refers to how easy or difficult it is for the reader to understand and understand the content of a written essay. Good readability makes it easier for readers to interact with the writing and helps keep their attention. On the other hand, poor readability can lead to confusion, frustration, and even loss of interest in the material. Several factors contribute to the readability of a written document, including the use of simple and clear language, the organization and structure of the text, and the format and layout of the writing. One of the most important aspects of readability is the use of clear, simple language. Using language that is complicated or full of jargon can make it difficult for the reader to understand the content and can even lead to frustration or confusion. By using simple, direct language, students can ensure that their message is conveyed effectively and clearly to the reader. The organization and

structure of the text also plays an important role in readability. A well-organized document with a clear structure helps readers follow the logical flow of content and makes it easier for them to understand the main points raised. Conversely, a poorly organized document can be confusing and difficult to follow, leading to a loss of interest in the document.

Style trait in writing is important because it allows students to express their individuality and creativity. It sets the student writing apart from others and gives their work a unique voice and perspective. Developing a personal writing style takes time and practice, as it involves finding their own voice and learning how to effectively communicate their thoughts and ideas in writing. It involves finding a balance between being honest with their own beliefs and values and being able to connect with and engage an audience. In addition to helping writers express themselves and stand out, a personal writing style can also contribute to the overall quality of an essay. A student with a strong sense of their personal style can write more coherent and cohesive text because they have a clear understanding of their own tone of voice and how to use it effectively.

Grammar is an essential aspect of writing that allows students to effectively communicate their ideas and thoughts through language. It's the set of rules that govern the structure and usage of words in sentences and help ensure that the writing is clear, concise, and easy to understand. One of the main benefits of good grammar is that it makes the writing more effective. When students use correct grammar, they can communicate our ideas logically and coherently, making it easier for the reader to follow their thought process. This is especially important when trying to convey complex or abstract ideas, as it helps ensure that readers can fully grasp the concepts they are trying to convey. Another important aspect of grammar is that it helps establish credibility. When students use wrong grammar, it can create a negative impression on the reader and make them not take us seriously. On the other hand, when they use correct grammar, it shows that they have put effort into our writing and are serious about communicating our ideas effectively. Grammar is also important for maintaining consistency in writing. By following rules of grammar, students can ensure that their writing has a consistent style and structure, making it easier for readers to follow.

3.6 Methodology

3.6.1. Feedback Formula

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 [19].

3.6.2. Overall Models

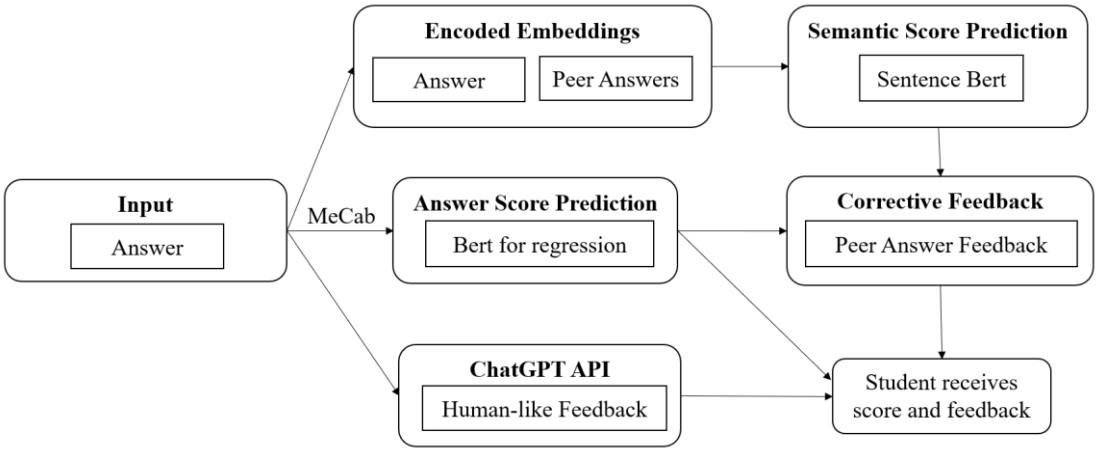

Figure 11: Overall AEF models

In figure 11, the proposed system will generate the peer answer feedback from the answer score and semantic score. It will also use the human-like feedback from another source, ChatGPT API. By providing the students will these two types of feedback in our system, we can evaluate which type of feedback is preferable by the students.

In details, the answer needs to be segmented by the morphological analyzer first before it is input into the answer score prediction model. The answer score prediction component then employs a Bert Neural Network to perform a regression task and generate a score for the answer, which will also be used as a metric for generating corrective feedback. The system will also encode the student's answer and other peer answers as embeddings. The reason for this step is we will do a semantic search to find the related peer answer. The output of it is the semantic score. The semantic score will finally be combined with the answer score from the previous step to generate the corrective feedback. Along with the corrective feedback, the human-like feedback and the score are the 3 components that will be given to the student.

3.6.3. Corrective Models

Corrective models have two tasks: predicting the score of the student’s writing (score pre-diction 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:

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

Peer Answer A	AS	SS
「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。。。	4	0.0573

Peer Answer B	AS	SS
西洋文化の基底には、自分の考えに相手を同意させる。。。	5	0.0767

Peer Answer C	AS	SS
西洋 人 は 基本 的 に 他人 は 自分 と は 異なる 人間 と 見な。。。	10	0.5261

Figure 12: Example of corrective feedback

In the Figure 12. we can see a list of 3 possible feedback that the system can choose as the feedback. Let's not regard the text answer, we only need to know that the student 's answer score is predicted to 3 by the system, so now the system will do a search of all peer answer that have the score between 3 and 6. At first, answer A and B were selected because their answer score fall within the range of 3 to 6. Then, because answer A has a closer semantic score, it is chosen over answer B. The answer A will then be used as the corrective feedback.

3.6.4. Score Prediction with Machine Learning

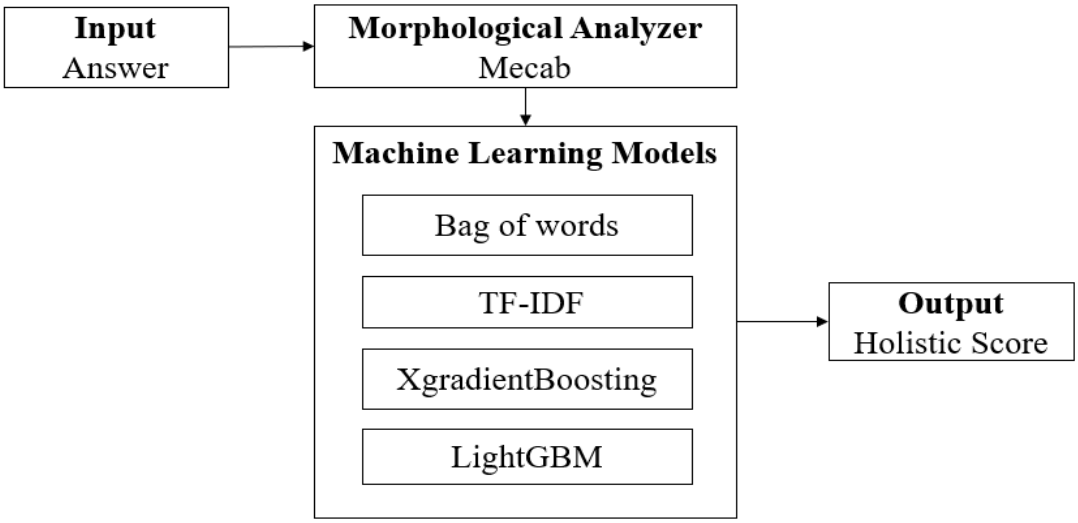

Figure 13: Score prediction model using Machine learning methods

For the score prediction task, we experiment with a list of machine learning methods like Bag of words, TF-IDF, XgradientBoosting and LightGBM. Figure 13 shows an overview of how we calculate the score based on the student answer. After the experiments, we pick the combinations that yield the highest accuracy among them. In these models, we only use a few training features like answer and the overall score. The Bag of words model is a representation of text that describes the occurrence of words within a document. It involves two things: a vocabulary of known words, and a measure of the presence of known words. It is called a "bag" of words because any information about the order or structure of words in the document is discarded. But This model is only concerned with whether a known word occurs in a document, not where in the document. TF-IDF (short for Term Frequency-Inverse Document Frequency) is a numerical statistic intended to reflect how important a word is to documents in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases linearly with how often the word appears in the document but is offset by the frequency of the word in the corpus, which can help adjust for the fact that some words are more common than others. XgradientBoosting is a machine learning algorithm that uses gradient boosting to make predictions. Gradient boosting is an ensemble learning technique that combines the predictions of multiple weak models to create a strong and accurate model. XgradientBoosting is an efficient and scalable gradient boosting implementation. One of the main drawbacks of using XgradientBoosting is that it can be

computationally expensive, especially for large data sets. XgradientBoosting is a gradient boosting implementation, so it creates many weak models and combines them into one strong model. This means that training can take a long time, especially if the dataset is large and the number of weak models is large. Additionally, XgradientBoosting may overfit the training data, which can lead to poor performance on test data. This can be addressed using regularization techniques, but it can still be an issue for some datasets. LightGBM is also a gradient boosting framework that uses tree-based learning algorithms. It was designed with efficiency and scalability in mind, making it one of the most popular gradient enhancement algorithms in machine learning. LightGBM is a type of gradient boosting that uses a histogram-based algorithm to make predictions, which can run faster and use less memory than other gradient boosting implementations. It is commonly used for classification and regression tasks in various fields. By combining these machine learning models, the system can predict a student's overall score using only that student's answer as input.

3.6.5. Score Prediction with Neural Network

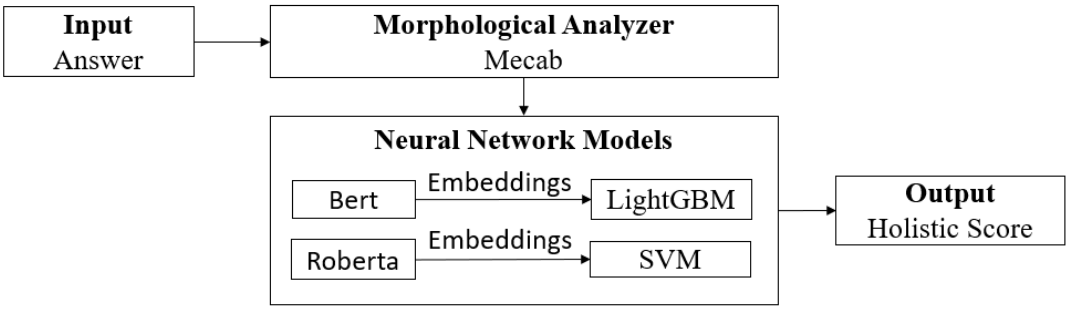

Figure 14: Score prediction model using neural network

In this experiment, instead of using the Bag-of-words or TF-IDF model, the Bert model was used to extract the contextual meaning of the answer. Figure 14 has the same input and output of the machine learning score prediction model, but the use of model is different. 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 de-fault morphological analyzer, and with the same number of training features.

3.6.6. Feedback Generation

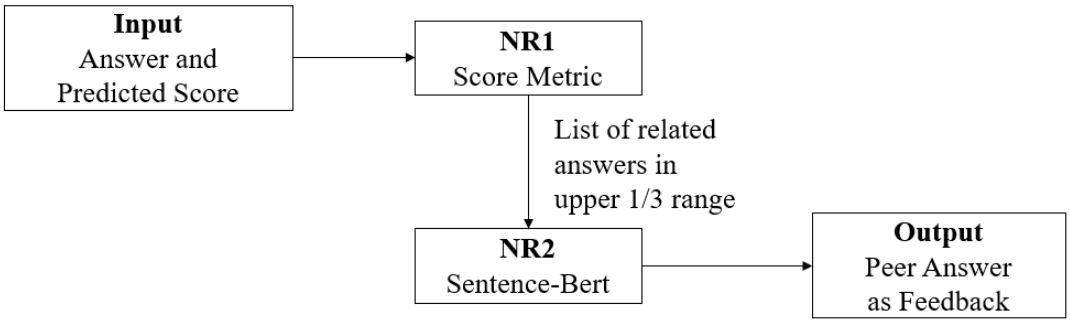

Figure 15: Feedback generation model

Figure 15 shows the use of answer score metric (NR1) and sentence similarities metric (NR2). Our 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 we use the Japanese Sentence Bert Japanese ("sonoisa/sentence-bert-base-ja-mean-tokens") model [20] 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.

Chapter 4

Experimentation

4.1 System Experiment

Our proposed system in previous chapter will generate the corrective peer-answer feedback. After the implementation, a new question arise, is the feedback can actually be used by the student? Instead of measuring the usability of the peer-answer feedback, we will try to measure the difference between one type of feedback to another. In this case, we want to evaluate how the student thinks about the peer-answer feedback, and the human-like feedback. By doing the experiment, it can help us shed the light on what the type of feedback that the student think they might need to improve their score.

4.1.1. Score design

Step 1. Preprocessing. Steps such as stopwords, punctuation, and morphological analysis are applied to student answer before it is fed into a model. Stopwords are removed from student answer because they add noise to the data and make it more difficult for the model to extract useful information. Punctuation marks can also introduce noise into data, making it more difficult for the model to extract useful information. Removing punctuation helps simplify the student answer and make it easier to process for the model. Morphological analysis step break words down into their constituent parts to help the model understand the data underlying structure. This step is necessary in our score prediction task because it can aid in determining the underlying meaning of Japanese words even when they are used in multiple forms.

Step 2. Machine learning models. We use these 4 models are the baseline to build our score prediction model: LightGBM, XgradientBoosting, Bag of Words, and TFIDF. The features to be used in these models extracted are from the Python sklearn library. The library helps us to convert a collection of student answer to a matrix of token counts. Also, with the use of StandardScaler function to standardize the data. In our experiments, we use our parameters as follows: test_size=0.25 and random_state=0. Finally, we use the mean square error function to calculate the difference between the prediction and test data.

Step 3. Neural network models. We used the Japanese pre-trained BERT and Roberta from Tohoku in our experiments because they are well-suited for text classification tasks due to their ability to understand the context of words in a sentence. This is important in

text classification because the meaning of words can change depending on the context. The architecture of these two transformers allows it to attend to different parts of the input, which aids in understanding the relationships between words and the overall meaning of a sentence.

These followings finetuning parameters are used in our neural network model.

```
TRAIN_BATCH_SIZE = 32
EVAL_BATCH_SIZE = 8
PREDICT_BATCH_SIZE = 8
LEARNING_RATE = 2e-5
NUM_TRAIN_EPOCHS = 3.0
MAX_SEQ_LENGTH = 128
WARMUP_PROPORTION = 0.1
SAVE_CHECKPOINTS_STEPS = 1000
SAVE_SUMMARY_STEPS = 500
```

4.1.2. Feedback design

We create our own feedback model from the SentenceBertJapanese("sonoisa/sentence-bert-base-ja-mean-tokens") [20] and finetune it to our needs. Our feedback model uses the BertJapaneseTokenizer to convert raw text into a format that can be understood by the Sentence Bert Japanese model. Then we calculate the mean pooling and encode all the embeddings. We use the test_size = 0.2 for train dataset and test_size = 0.5 for test dataset. In this model, we focus only on the “Char” feature which contain all the important words in the student answer.

4.2 Educational Experiment

The preliminary experiment included 2 international students with high Japanese skill (n=2, female=0, male=2). They are JAIST students pursuing a master and PhD's degree. The experiment was designed as a writing task problem using the questions from the Riken Dataset. The entire description and procedure for participants to follow were made available online. And the participants followed those directions to complete the experiment.

Step 1. Introduction. This section explained the purpose of the experiment as well as the expected behaviors of the participants when they joined the experiment. The following was written in the introduction:

"You will answer 2 questions within a limited given time. This experiment takes approximately about 30 minutes, with 15 minutes for each question. The only thing that matters is to complete the question since it is necessary for the verification and validation of the hypothesis model."

Step 2. Experiment procedure. This procedure outlined the steps that participants should take to achieve a positive outcome from answering the experiment's questions. The following procedure was written:

1. *Read the QUESTION section to ensure that you understand the task clearly.*
2. *Read the FEEDBACK section to evaluate your writing.*
3. *Do the survey in QUESTIONNAIRE section.*

Step 3. QUESTION. This part includes the first 3 questions extracted from the Riken Dataset that the participants must answer:

1. Question 1: 「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。
2. Question 2: 「世界と私の間に言葉の橋を架けることはできるけれど、そのおかげで世界と私の間 に橋が架かってしまうので、距離も生まれてしまう」とあるが、これはどういうことか。本文に 即して句読点とも七〇字以内で説明せよ。

Section 4. Feedback. Students will receive two types of feedback. The first feedback type is the peer answer, and the second feedback type is the human-like feedback generated from GPT-3 models.

Section 5. QUESTIONAIRE. This section included a quiz that asked participants to reflect on their progress through the question to distill their achievement from the task. The student will answer this questionnaire two times, first time is for the peer answer feedback, and second time is for the human-like feedback. The detail of the questionnaire can be seen in Table 2.

Table 2: Questionnaire for the feedback

No	Quiz	Answer
1	I am a student of	Elementary / Undergraduate / Graduate
2	Are you a Japanese native speaker	No / Yes
3	Feedback is appropriate for the question	1 / 2 / 3 / 4 / 5
4	Feedback would be a welcome addition to a lecture	1 / 2 / 3 / 4 / 5

5	Feedback makes the writing task more interesting	1 / 2 / 3 / 4 / 5
6	Feedback has potential to enhance writing ability	1 / 2 / 3 / 4 / 5
7	Feedback gives me insight to understand the question	1 / 2 / 3 / 4 / 5
8	Feedback helps me better understand my strength and weakness in writing	1 / 2 / 3 / 4 / 5
9	Feedback helped me more engaged in my writing task	1 / 2 / 3 / 4 / 5
10	Feedback will have bad effect on student writing	1 / 2 / 3 / 4 / 5
11	Feedback distracted me from my writing task	1 / 2 / 3 / 4 / 5
12	It is not fair for students that do not have access to the feedback	1 / 2 / 3 / 4 / 5
13	Please give us suggestions on how we can improve the feedback	Text

Chapter 5
Evaluation

5.1 System Evaluation

5.1.1. Preliminary Results for Score Prediction

Score Prediction: Four machine learning 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. We also experimented with the neural network models but they did not perform as well as the four machine learning methods. The highest accuracy we could get for the neural network model is 0.65.

As can be seen from Table 3, LightGBM and Bag of words achieve the highest accuracy (0.746) among other machine learning methods. In Table 4, comparing the neural network models from our proposed method and Mizumoto and Takano, we can see that our neural network model performs with a low accuracy score. The total of 14 questions from the dataset even though are balanced, and well annotated on 13 questions, but from the fine-tuning process of neural network, hyper-parameters settings might have contributed to the low accuracy and underfitting.

Table 3: Score prediction accuracy

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 4: Our neural network accuracy comparing to other models.

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

5.1.2. Preliminary Results Feedback Generation

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. So instead of measuring the accuracy or RMSE metric like score prediction task. We will try to measure the feedback based on subjective measurements on how it is useful for the student writing. We will also compare the generated feedback based on peer answer with the human-like feedback from the GPT-3 model. The educational evaluation of feedback will focus on the two questions from the Riken Dataset. Question 1: 「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 2: 「世界と私の間に言葉の橋を架けることはできるけれど、そのおかげで世界と私の間 に橋が架かってしまうので、距離も生まれてしまう」とあるが、これはどういうことか。本文に 即して句読点とも七〇字以内で説明せよ。

One downside of this approach is we are encoding all the peer answer as the embeddings, so when we calculate the Semantic Score and finding the similar answers to be used as feedback, the task is very CPU intensive, which can result in very long wait time. In our research, we use an Intel Xeon CPU 1 core, 2.30GHz which will take about 2 minutes to process 1 question with 10 000 rows of data. Even though the peer answer feedback method might be useful in term of education, from system point of view, the lookup search using Sentence Bert is not very effective and will take lot of time to process many questions. The worst-case scenario is for all the 10 questions in the corpus, it would take 20 minutes for the student to get the feedback. We also try to use better CPU in our research, but the wait time is still noticeable by the student. Figure 16 is an example of how the feedback was chosen for Question 1, with the following query.

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 16: Example of feedback generation

5.2 Educational Evaluation

5.2.1. Generated Feedback for participants

Table 5, and 6 below has the 4 following features: student number, student answer, human-like feedback, peer-answer as feedback. We only include one student answer as an example so we can have an overview of how a student approaches the questions, write an answer, and receive the two types of feedback.

Question 1: 「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。

Table 5: Feedback for question 1

Student Number	Student Answer	Human-like Feedback	Peer Answer as Feedback
1	緊張は不安と疑惑につながり、疑念は偏見につながり、人々は自分の快適さと利益を確保しようとし続け、西洋文化を騒がせます。	この文章は文法的に正しく、使用されている語彙も適切です。しかし、緊張が不安や疑惑に繋がること、疑念が偏見に繋がることは明確ですが、最後の文 "人々は自分の快適さと利益を確保しようとし続け、西洋文化を騒がせます"は文脈からは明瞭ではないです。 ライターが伝えようとしていることについてもう少し詳細を提供することは有効でしょう。文章はより詳細で、文脈に沿った情報を持ったり、最後の文を再構成することでより効果的になるかもしれません。	西洋は民族や言語や文化を異にする多くの国々が集まっている為、基本的に他人は異人と見なし自分の考えに他人を同意させる必要があるということ。

Question 2: 「世界と私の間に言葉の橋を架けることはできるけれど、そのおかげで世界と私の間 に橋が架かってしまうので、距離も生まれてしまう」とあるが、これはどういうことか。本文に 即して句読点とも七〇字以内で説明せよ。

Table 6: Feedback for question 2

Student Number	Student Answer	Human-like Feedback	Peer Answer as Feedback
1	社会や人から与えられた情報を処理する感覚がない場合、言語は、人々が社会や人々が表現したいことを理解するのに役立つ非常に効果的な媒体として使用できますが、同時にノイズも追加します。元の情報へのノイズ情報エントロピー。	この文章は文法的には正しいですが、意味は少し曖昧です。「社会や人から与えられた情報を処理する感覚がない場合」という部分から、「言語は、人々が社会や人々が表現したいことを理解するのに役立つ非常に効果的な媒体として使用できますが、同時にノイズも追加します」という部分にかけて、より明確な解釈や文章の構成をすることで、意図をより明確に伝えることができるかもしれません。	自ら他者に理解される言語を話せるようになることは、自分を世界に発信する利点もあるが、世界と自分との一体感を失ってしまうということ。

5.2.2. Questionnaire from participants

Table 7 is the summarized results of the questionnaire from the two participants. Even though the number of participants is limited, but there's an indication that the students prefer the human-like feedback more. "Feedback gives me insight to understand the question", "Feedback helps me better understand my strength and weakness in writing" are subjective questions but they are correlated to the suggestions we received from the students. Some suggestions from the participants are "The determination of whether the context statement is really relevant needs to be improved.", or "Please give more concrete examples". From those we can roughly conclude that the feedback can be a nice add-on to the student writing, but both the human-like and peer-answer feedback needs to be more detailed and have better explanation. Other questions like "Feedback will have bad effect on student writing", "It is not fair for students that do not have access to the feedback" are both being disagree by the participants.

Table 7: Questionnaire for human-like and peer answer feedback

(1 is completely disagree, 5 is completely agree)

Question	**Human-like Feedback**	**Peer Answer as Feedback**
Feedback is appropriate for the question	4	3
Feedback would be a welcome addition to a lecture	3.5	3
Feedback makes the writing task more interesting	3	3.5
Feedback has potential to enhance writing ability	4	3.5
Feedback gives me insight to understand the question	4	2
Feedback helps me better understand my strength and weakness in writing	4	3
Feedback helped me more engaged in my writing	3	4

task		
Feedback will have bad effect on student writing	2	2.5
Feedback distracted me from my writing task	1.5	2
It is not fair for students that do not have access to the feedback	2.5	2

Chapter 6
Conclusion

In this research, we applied the writing theory, which is mostly used in the US, and make adjustments so it could fit the Japanese intermediate education. We also built the first Japanese Automated Feedback System that can generate a feedback based on the student's answer and other peer answers. From qualitative research, we reached a conclusion that an automated feedback system needs to cover at least one type of feedback from this list: structure, readability, style, word, convention and holistic. This answers our first research question "What feedback to include in Japanese automated essay feedback system?", but more qualitive methods should be done to justify this.

Our first assumption is that holistic feedback is crucial for the success of the student when learning to write an essay. But experiments indicate that students might prefer to use feedback that is more expressive like the written feedback from a teacher, instead of just showing related peer answer to them. Even though the number of participants is limited, the results show that the students are more positive towards the auto-generated human-like feedback from GPT-3. Our proposed method which uses related peer answer as feedback even though in theory, might help improve the overall score of the writing, but the lack of context and explanation can result in confusion for the students. Students also find that the feedback in both methods are helpful for their writings, but because we do not conduct a writing test, there is no concrete evidence that the feedback can improve the student writing or not. Overall, our research is not able to fully answer the second research question "What feedback improve writing of Japanese students in intermediate education".

For the last research question "What method can improve score prediction accuracy of an essay?", we experimented with four machine learning methods and two neural network models but the highest accuracy we could get is 0.746. Our neural networks might have suffered from underfitting and bad hyper-parameters tuning, which results in unexpected low accuracy. From the context of feedback generation, the score is an important metric to indicate which feedback will be chosen by the system. If we can improve the score prediction accuracy, it would contribute to better generated feedback.

A limitation in this research is that the lack of pair-wise format dataset of teacher feedback. Our proposed methods can be applied for any text in a pair format using the score as the relationship indicators, but the lack of dataset make it difficult to implement

an automated feedback based on teacher knowledge. The need of labelling is also another big factor when building a feedback system because our proposed methods mostly use the supervised learning techniques, but the education domain exists many unlabeled essays.

In future works, we will evaluate thoroughly the feedback with more participants, then build a new model that can show the relationship between all the traits in the 6+1 writing theory. We will also work on finding the relationship between the automated feedback system and the real-life students who have studied the writing theory in class.

Publication List

Huy Phan, Shinobu Hasegawa, and Wen Gu, Implementation of Automated Feedback System for Japanese Essays in Intermediate Education, received Kunifuji Award in the 17th International Conference on Knowledge, Information and Creativity Support Systems (KICSS 2022). (Accepted and presented)

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] 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.

[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] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Oct. 2018.

[12] “https://github.com/cl-tohoku/bert-japanese.”

[13] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” Jul. 2019.

[14] “https://huggingface.co/cl-tohoku/roberta-base-japanese.”

[15] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Aug. 2019.

[16] L. Floridi and M. Chiriatti, “GPT-3: Its Nature, Scope, Limits, and Consequences,” Minds Mach (Dordr), vol. 30, no. 4, pp. 681–694, Dec. 2020, doi: 10.1007/s11023-020-09548-1.

[17] “https://openai.com/api/.”

[18] “https://github.com/Hasegawa-lab-JAIST/huyphan-6-writing-trait-feedback.”

[19] “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.”

[20] “https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens.”