International Journal of Learning Technologies and Learning Environments
International Institute of Applied Informatics
2025, Vol. 8, No. 1, IJLTLE844

Correction Method for Character Recognition of Handwritten Answers in Multiple Formats for Short Essay e-Learning Systems

Takahiro Yamasaki [*] , Ayako Hiramatsu [†]

Abstract

We are aiming to develop an e-learning system for improving Japanese language proficiency. This system not only provides instant scoring and advice in response to learners' answers but also automatically generates questions to offer a wide range of problems across various fields. In this paper, as the first step for this purpose, we aim to accurately extract the content described in handwritten essays. When dealing with handwritten characters, sufficient accuracy cannot be obtained with OCR software alone due to noise, distortion, and idiosyncratic handwriting. Therefore, we first train a neural network for character recognition using a handwritten character database, and then divide the essay images into single characters to determine the most likely character candidates. Furthermore, we treat character identification as a fill-in-the-blank problem for sentences and use BERT's Masked Language Model task to determine characters that form natural sentences. Applying this method to handwritten short essays written in multiple formats has enabled more accurate character extraction than before.

Keywords: e-learning, handwritten character recognition, Japanese essay summarization, natural language processing

1 Introduction

Currently, e-learning, which utilizes information and communication technology (ICT) in various situations, is being used in education and learning. MOOC (Massive Open Online Course) platforms have been established as on-demand e-learning, and many universities have joined in. In MOOCs, course content can be accessed through the internet, and in addition to attending lecture videos, online discussion boards for questions and opinions exchange, as well as learning evaluations through class assignments are provided [1]. In such on-demand e-learning, creating educational content and understanding learners' situations are essential, and the complexity of content creation is a challenge for providers in actual operation. For learning evaluations in class assignments, multiple-choice questions and automatically graded word inputs are often used for providing immediate feedback.

[*] Dept. of Electrical, Electronic and Information Engineering, Osaka Sangyo University, Osaka, Japan
[†] Dept. of Information Systems Engineering, Osaka Sangyo University, Osaka, Japan

Descriptive assignments are often graded based on the presence or absence of keywords, and immediate evaluation is difficult for highly flexible descriptive assignments.

Furthermore, not only in high schools but also in universities, Japanese language proficiency is emphasized as part of first-year education. It is said that to improve writing skills, one must understand the basic rules and practice writing many sentences. E-learning, as a self-learning environment for written texts, enables effective learning by using limited time efficiently and solving problems according to one's level. With the recent spread of online classes, the demand has been high, and many universities have reported its adoption. By identifying problem points and considering corrective methods, learning can be more effective. As a self-learning environment, e-learning allows for effective learning by making efficient use of limited time and solving problems tailored to one's level. The demand for e-learning has been high lately, and many universities have reported its adoption due to the widespread use of online classes.

However, for an assignment of summarizing a long Japanese essay, after submitting the answers, the grader reads the text and provides advice by writing comments, which are then sent back to the learner. It is not possible to provide real-time advice, and there is a significant time lag before learners can consider corrections, preventing them from tackling more assignments. Additionally, in e-learning, answers are entered using a PC or tablet, so the use of auxiliary functions for text proofreading and typo corrections is assumed. As a result, even in descriptive assignments, the fundamental practice of writing characters by oneself is not performed.

In this study, we aim to develop an e-learning system capable of self-learning by focusing on summary exercises of short essays as Japanese writing practice [2]. The following items are important in this learning system:

- A function to recognize handwritten answers
- A function to immediately grade and evaluate the written summary and provide feedback on improvements
- A function to automatically generate educational content, enabling learners to tackle a large number of problems

In this paper, we first discuss the objectives and challenges of the short essay summarization problem and describe the overview of the e-learning system that we aim to develop. Additionally, this system is based on handwritten short essays. With the goal of accurately reading handwritten characters written on manuscript paper, and we propose a two-step method: (1) handwritten character recognition using machine learning models, and (2) character correction through sentence estimation, to accurately comprehend the content. Furthermore, for future system development, we investigate and compare whether characters can be recognized with equivalent accuracy in different writing formats, such as vertical writing, horizontal writing, and free format, to consider improvements.

2 Guidelines for a Essay Summary Problem

To improve the ability to write essays and other compositions, it is said that one should understand the basic rules and then practice writing repeatedly. There are two types of essay problems in the format of given prompts: expressing one's own opinion and summarizing a given text. This study focuses on summary problems, which require more fundamental

practice, emphasizing reading comprehension to understand the presented text and the ability to concisely express its content. To implement summary test learning in e-learning, it is necessary to provide a variety of subject texts to be summarized and to evaluate answers promptly, giving feedback on areas for improvement.

There are various approaches to automated essay scoring, including those in practical use [3]. In particular, research on the automatic evaluation and scoring of English essays has a long history [4][5][6][7], with scores being given based on keywords, grammatical correctness, phrasing, and structure. For assessing logical structure, cue words (connective expressions) and rules based on phrasing in the text are used. Recently, methods utilizing artificial intelligence and machine learning technologies have also begun to be proposed [8][9]. These essay test assume opinion expression and emphasize logical development. In some cases, a large number of essays scored by experts in advance may be required, but using a large number of past response examples is not feasible for providing diverse topics.

Whereas existing systems can be used to evaluate grammatical correctness, from the perspective of summary problems, it is crucial to concisely summarize without deviating from the main theme of the task text, and the evaluation of the description content becomes more important than logical development. As for the descriptive content, it is necessary to extract essential parts from the task text, but simply quoting the text does not lead to practice in Japanese expression based on one's own reading comprehension; concise paraphrasing is required. In addition, there are various ways to summarize, and multiple correct examples can be considered based on the use of keywords and the order of explanation.

In the summary problems considered in this study, we roughly divide the scoring criteria from the perspective of content, rather than grammar or writing rules, into four levels:

A: The main theme is concisely summarized (correct answer).

B: Not deviating from the main theme, but with some excess or deficiency.

C: Excerpt from the task text, but deviating from the main theme.

D: Contains content outside the task text, no longer a summary.

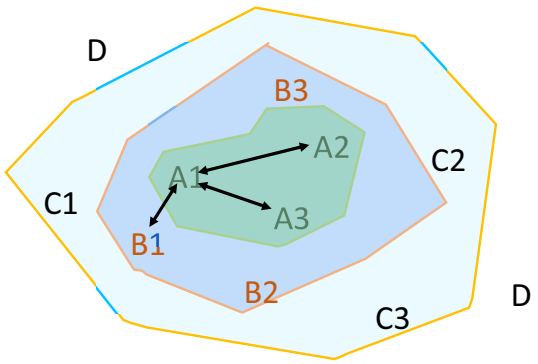

Figure 1: Distribution Image of Evaluating Essay Summary

Figure 1 shows the distribution image of these responses. The evaluation boundaries are ambiguous and may vary depending on the scorer. Therefore, in actual evaluations,

multiple scorers often assess the responses. Additionally, since there is no single way to express ideas concisely, evaluations can differ even if similar words are used when comparing words alone. Given these factors, it is necessary to determine evaluations by comparing multiple correct examples. However, when giving comments, it is essential to advise on the differences to approach the nearest correct example. For instance, in Figure 1, it is appropriate to give modification advice for response B1 to approach A1 rather than to move towards the center of the distribution of correct examples.

3 Short Essay Self-learning Systems

For self-learning of the aforementioned summarization problems, the proposed e-learning system takes an approach that evaluates the answers based on their similarity to multiple correct examples and provides comments on the differences between the answer and the correct examples. Additionally, the system includes an input interface for actually writing the answers. Figure 2 shows an overview of the proposed system. This e-learning system has the following functions:

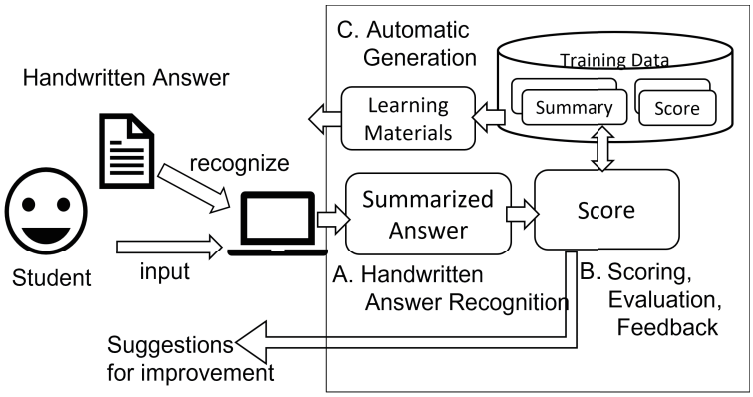

Figure 2: Overview of the Short Essay Self-learning Systems

3.1 Handwritten Answer Recognition Feature

Handwritten character recognition can be achieved by utilizing traditional machine learning models for character image recognition, which are designed to recognize individual characters. However, when dealing with a collection of characters written as a sentence, character boundaries can become unclear, leading to misrecognition. In this system, since the characters are written on manuscript paper, it is possible to handle this issue by performing individual character segmentation based on the gridlines.

Moreover, as the text involves summarized sentences, the types of characters that appear can be predicted and corrected from the training data. By leveraging these techniques, the e-learning system can accurately recognize handwritten characters in a sentence context and integrate them into the evaluation and feedback process.

3.2 Instant Scoring, Evaluation, and Feedback Function for Written Summary

To automate the evaluation process, multiple correct examples are prepared in advance, and methods for calculating the similarity between them are examined. For the problem text, multiple correct examples, and answer text, they are represented as graphs with words as nodes and connections between words as edges. In order to consider the similarity between the graphs, the edit distance is calculated. The difference in the words used is corrected by the similarity between the words in the edit distance[10].

By using this approach, the e-learning system can efficiently evaluate students' written summaries by comparing them to multiple correct examples based on the calculated similarity. This method takes into account both the words used and their relationships, resulting in a more accurate and comprehensive evaluation of the summary's quality.

3.3 Automatic Generation of Learning Materials Feature

It is possible to sequentially add new problem texts by extracting appropriate-length candidates from news articles, essays, and other content published on the web. On the other hand, it is necessary to prepare multiple correct examples of summaries for these texts. Various automatic summarization techniques have been proposed for this purpose.

Furthermore, it is possible to generate multiple correct examples by automatically creating different expressions based on a single correct example. However, the correct examples needed for this system's evaluation approach are summaries used for similarity comparison, and they do not necessarily have to be complete sentences. Therefore, we are considering the automatic generation of evaluation-oriented correct examples.

By utilizing automatic summarization and paraphrasing techniques, the system can generate multiple evaluation-oriented correct examples, enhancing the effectiveness of its evaluation approach.

In this paper, we propose a feature to recognize handwritten answers, which is the first part of the short essay self-learning system. This functionality is not necessary when students directly input their answers into a computer. However, since the goal of this research is to facilitate repetitive learning through handwriting, it is essential to accurately read the characters of the answers. In the next chapter, we will demonstrate character recognition techniques using machine learning and natural language processing.

4 Character Recognition from Handwritten Essay Answers

4.1 Outline

To grade handwritten short essay responses, they first need to be converted to electronic data. However, commercial OCR software cannot read them due to noise and distortion. Prior to 2018, methods using N-grams were proposed. This involved dividing the text into n-character segments, shifting one character at a time, and calculating the probability of each n-character combination occurring. A method exists that assigns a score of -1 to all n-characters with a probability below a certain threshold, and detects incorrect characters from the total score [11]. Since the release of BERT in 2018, methods utilizing BERT have become mainstream. Techniques such as masking a portion of the text and predicting it have been proposed to improve recognition accuracy [12][13].

Language models were traditionally formulated as the sum of log-likelihoods of the conditional probabilities of predicting the next word given the words output so far. However, such autoregressive models are not suitable for masked language models (MLMs) like BERT, which predict words bidirectionally. A Pseudo Log-Likelihood (PLL) score has been proposed as a mechanism for scoring the naturalness of sentences in MLM models [14][15]. Instead of calculating probabilities sequentially from the beginning of the sentence, this method uses the sum of the log-likelihoods of the conditional probabilities when predicting words masked with MASK as a score representing naturalness. It has been shown that this method can determine the correctness of sentences with similar or higher accuracy than language model scores in autoregressive models.

In this paper, we perform character recognition after distortion correction and noise removal in handwritten short essays, and set the characters with high probability as candidate characters. We then use a method to score the naturalness of the sentences for these candidate characters and correct the answers by selecting the characters determined to be the most natural. Figure 3 shows an overview of this process.

In this idea, there is a possibility that typographical errors, such as typos and missing words, will be corrected. However, the focus of this e-learning system is on summarizing problems. It emphasizes understanding the content of the essay in question, ensuring that no essential elements are missed, and whether a logical sentence can be written. Therefore, we are not targeting obvious errors such as typos in this system.

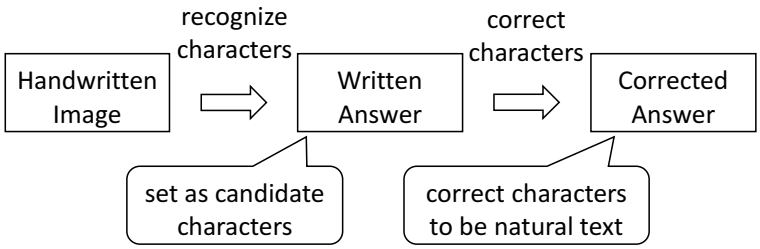

Figure 3: Outline of Character Recognition

4.2 Character Recognition from Images

Character recognition from handwritten manuscripts is carried out using the following procedure. First, the manuscript is scanned and imported as an image. After correcting for image distortion, character bounding boxes are detected, which allows the image to be divided into individual characters. Character recognition is then performed on each segmented image, one character at a time. A neural network (NN) is used for image recognition in this process. The input to the NN is a monochrome image (32x32 pixels), and the output consists of characters and symbols (3043 types). For NN training, images from the handwritten kanji database ETL9B[16] are used. The flow of character recognition is shown in Figure 4. In the example in Figure 4, the character with the highest probability is correctly estimated. However, the correct character is not always the one with the highest probability in all cases. In the example in Figure 5, the character in the center has the second-highest probability as the correct answer. Moreover, the neural network computes probabilities even for very small values, as shown in the examples on the right and left. For subsequent answer correction, candidate characters are set with probabilities of 0.01 or higher, up to a maximum

of 10 characters. This is effective in reducing the time required for subsequent character estimation.

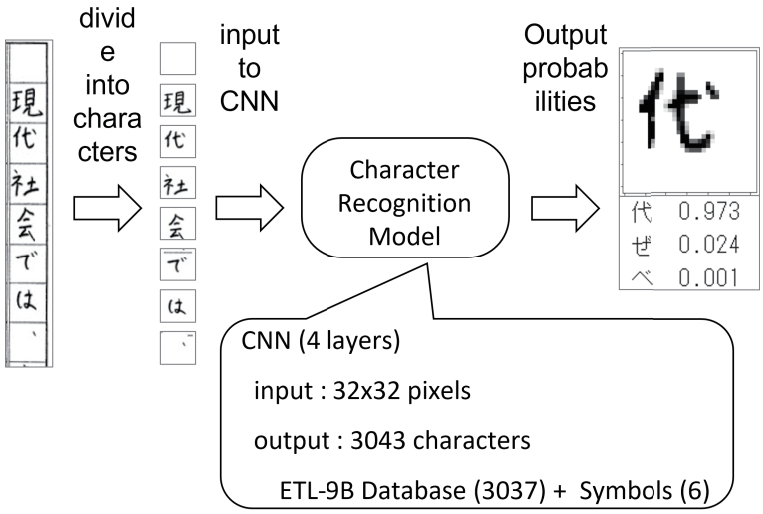

Figure 4: Character Recognition from Images

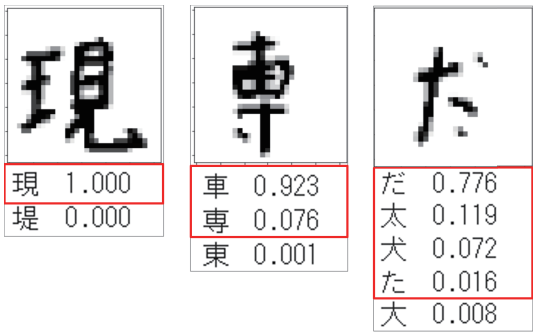

Figure 5: Setting Candidate Characters

4.3 Character Correction Based on Sentence Estimation

As shown in Figure 5, the character with the highest probability may not necessarily be the original handwritten character. Therefore, to perform answer correction, the step of determining the character is treated as a fill-in-the-blank problem for sentences. This task is handled in natural language processing, and a solution using BERT's Masked Language Model (MLM) is available. By using a mechanism that scores the naturalness of a sentence, the correction of the recognized characters from the images is performed by determining which candidate character is the most natural among those in the candidate list. The flow of character correction is shown in Figure 6. There are two types of candidate characters to fill in the □, and the naturalness of the sentence is scored for each case when the characters are applied. In the case of Figure 6, the sentence below is considered more natural, and

the candidate character below is selected. This is correctly determined. By repeatedly performing this character correction operation from the beginning of the sentence, the correct character can be selected even for characters with low estimation probabilities in character recognition, and the correct sentence can be restored.

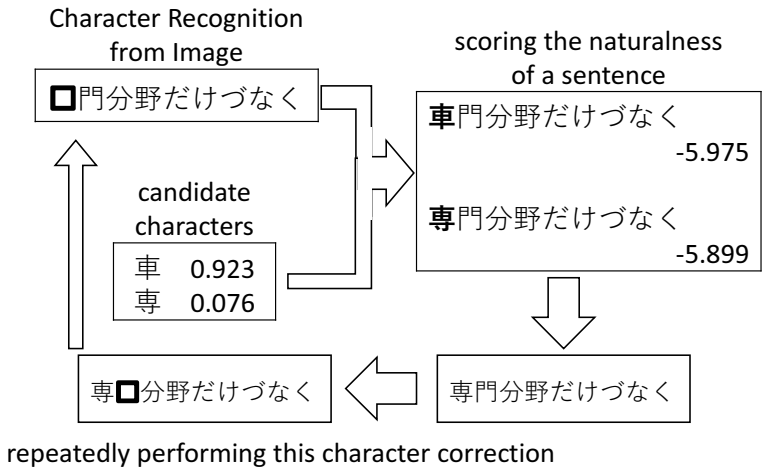

Figure 6: Character Correction Based on Sentence Estimation

5 Experiment Results

5.1 Effects of Character Correction through Sentence Prediction

We actually converted handwritten manuscript paper (400 characters) into images and performed character recognition and correction through sentence prediction. Figure 7 shows a part of the handwritten manuscript used in the experiment. For these manuscripts, we first carried out image segmentation, and then, we performed character image recognition alone, correction for all characters, and correction for characters with the first candidate character (the highest probability) less than 90%, 95%, and 98%. Table 1 shows the number of characters correctly restored for each case.

Table 1: Number of correct characters (Manuscript(a))

correction times	image recognition only	correct characters (number of correction target)			
		All (400)	less than 90% (150)	95% (187)	98% (255)
1st	287	294	293	298	294
2nd	—	300	300	311	299

With image recognition alone, only 287 characters were correctly identified, but by performing character correction for all characters, 294 characters were correct. Here, as

Figure 7: A part of the experimental manuscripts

shown in the left figure of Figure 5, in cases where the probability of the first candidate character is high, there is often no need to perform character correction again. Therefore, when the number of correction targets was changed based on the probability of the first candidate character, the best result was obtained when targeting those with a probability of 95% or less. This was also the case when corrections were made until the end of the sentence and a second correction was performed from the beginning. Furthermore, Table 2 shows the correction results for each manuscript. For the cases where all characters were targeted and those with a probability of 95% or less, the process was performed twice for each. As a result, it was found that performing character correction twice for characters with a probability of less than 95% yielded the best trend.

Table 2: Number of correct characters (Each manuscript)

manu script	image only	All characters		less than 95%	
		1st	2nd	1st	2nd
(a)	287	294	300	298	311
(b)	265	274	287	279	292
(c)	320	320	319	323	328
(d)	197	190	195	195	204
(e)	211	188	192	195	195

5.2 Effects of Character Correction through 2-gram Prediction

Japanese is a language with a close relationship between adjacent characters when targeting one character at a time. Therefore, it can be said that predicting sentences two characters at a time, rather than one character at a time, makes the judgment of the naturalness of the sentence smoother. Thus, as shown in Figure 8, we evaluate by scoring the sentences every two characters and calculating the most natural combination. The results are shown in Table 3. As a result, the number of correct characters increases, but there is a problem that the number of combination patterns becomes enormous.

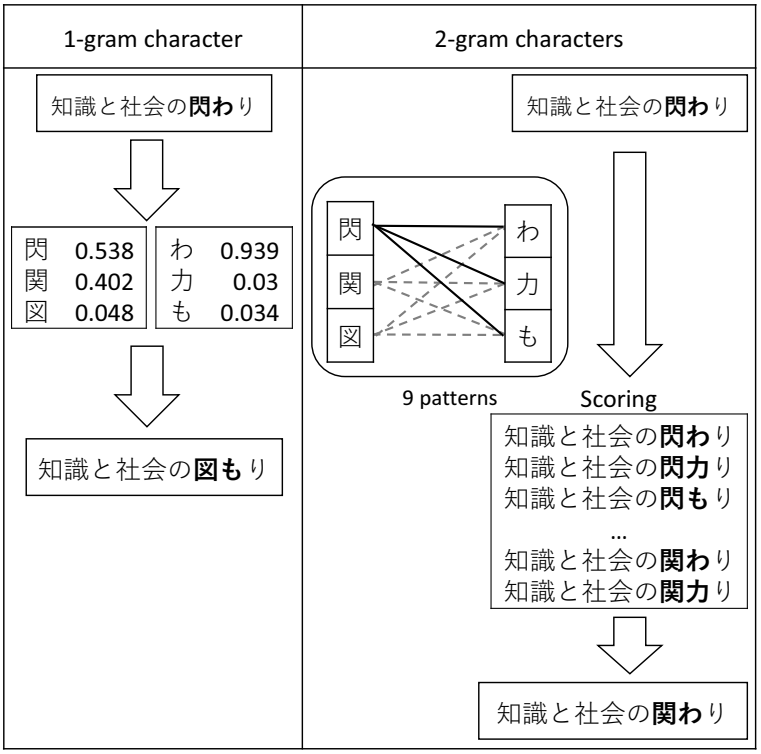

Figure 8: Character correction for every 2-gram characters

Table 3: Number of correct characters (for every 2-gram characters)

correction times	image only	1-gram 95%	2-gram 95%
1st	287	298	293
2nd	—	311	313

5.3 Recognition Results in Different Writing Formats

Until the previous section, the experimental subjects were manuscript papers divided by frames, and other writing formats were not addressed. In this section, as a preliminary phase for future system development, we investigate whether characters can be recognized with equivalent accuracy in different writing formats, such as vertical, horizontal, and free format writing. Figure 9 shows a part of the handwritten manuscripts used in the experiments. As before, we performed character correction twice, once targeting all characters and once targeting those with less than 95% confidence.

The results of applying this to horizontal manuscript paper (Manuscript (b)) and free format with only underlines (Manuscript (c)) are shown in Table 4. Since Manuscript (c) contains 373 characters, the values in Table 4 represent the proportion of characters that were correctly recognized. Manuscripts (a) and (b), which could be divided by manuscript paper frames in both vertical and horizontal writing, were recognized with reasonably high

accuracy. However, in the case of the free format, due to poor character separation, the recognition accuracy was comparatively lower.

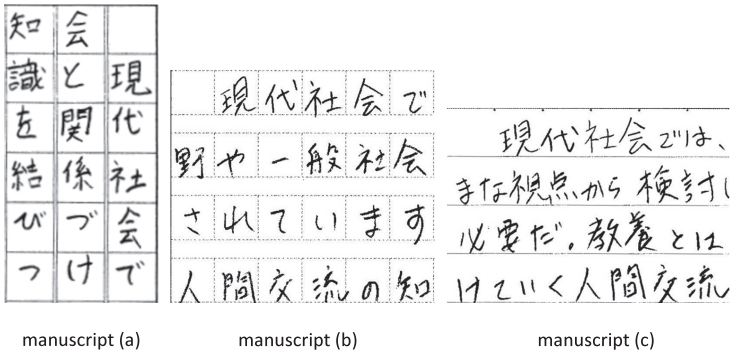

Figure 9: A part of the experimental manuscripts (Multiple Formats)

Table 4: Proportion of characters correctly recognized (Each Manuscript)

manu script	image only	All characters		less than 95%	
		1st	2nd	1st	2nd
(a)	71.8%	73.5%	75.0%	74.5%	77.8%
(b)	73.2%	74.0%	75.3%	74.8%	75.5%
(c)	40.1%	39.5%	39.8%	42.3%	44.1%

5.4 Character Correction through Improved Sentence Estimation Method

For Manuscript (c) in Figure 9, a part of the results from image segmentation is shown in Figure 10. With the current segmentation method, cases like character (c) with poor separation may lead to misrecognition. In the case of Figure 10, a three-character word is divided into four images, which cannot be correctly restored using the previous methods. Therefore, we improved the proposed method by adding 'no character' as an option among the candidate characters during character estimation. By adding this to the candidate characters as shown in Figure 11, it becomes possible to more accurately evaluate the naturalness of the sentence. Table 5 shows the results of the correction with this improvement. An increase in the proportion of correct characters was observed.

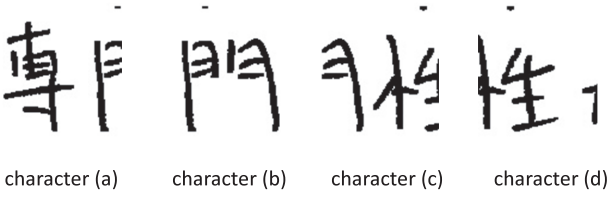

Figure 10: A Part of the Image Segmentation

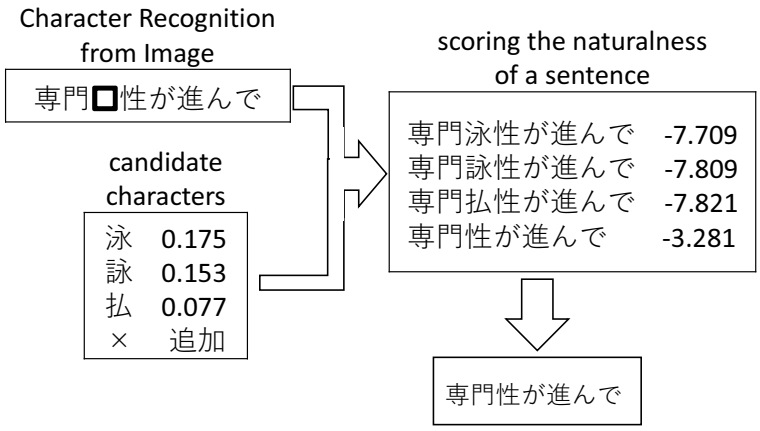

Figure 11: Improvement in Character Correction Based on Sentence Estimation

Table 5: Proportion of Characters Correctly Recognized (with 'No Character')

correction times	image only	previous method	No Characater
1st	40.1%	42.3%	45.3%
2nd	—	44.1%	48.2%

6 Conclusion

Currently, educational institutions are emphasizing Japanese language operational skills. To improve descriptive skills, it's important to repeatedly write after understanding the basic rules, and it's more effective to be pointed out inappropriate areas for learning. We aim to develop a descriptive answer scoring support system for improving writing skills. In this paper, we discuss the targeted summary format of short essays and describe the overall picture of the e-learning system we are developing. As the first step in this process, we aimed to accurately extract characters from handwritten manuscripts and proposed a method for correcting handwritten responses using BERT for sentence estimation. As a result, the number of characters correctly recognized increased compared to using image recognition alone. Furthermore, in writing formats where the separation of characters is clearly discernible, like manuscript paper, we confirmed that characters can be recognized relatively accurately in both vertical and horizontal writing. However, in the case of free format, due to issues with the accuracy of character separation, a significant drop in recognition accuracy was reconfirmed.

However, while the proposed method did increase the number of correct characters, it still did not achieve sufficient accuracy. We plan to explore methods to improve the character recognition model to accommodate smaller manuscripts and to enhance accuracy by expanding the dictionary and adding training based on the essays and questions given as short essay tasks. Additionally, we believe it is necessary to consider extending character recognition methods according to the writing format.

Moreover, in the automatic scoring of essays, typographical errors, omissions, and grammatical mistakes also become important evaluation criteria. The character estimation method proposed in this paper determines characters to form the correct words, which sometimes overlooks such errors made by learners. As an e-learning system, improving it to also point out such mistakes is one of our future challenges.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP23K02645.

References

[1] T. Fujimoto, Y. Ara, and Y. Yamauchi, "The Current Status of Learning Analytics Research for Massive Open Online Courses (MOOCs)," Japan Society for Educational Technology, Vol.41, No.3, 2018, pp.305–313.

[2] T. Yamasaki and A. Hiramatsu, "A Study of Correcting Handwritten Answers for Short Essay Self-learning Systems," Proc. 14th International Conf. on Learning Technologies and Learning Environments (LTLE 2023), 2023.

[3] T. Ishioka, "Latest Trends in Automated Essay Scoring and Evaluatio," Japanese Society for Artificial Intelligence, Vol.23, No.1, 2008, pp.17–24.

[4] J. Burstein and M. Wolska, "Toward evaluation of writing style: Finding overly repetitive word use in.student essays, " Proc. 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL '03), 2013, pp. 35–42.

[5] E. B. Page, "New computer grading of student prose, using modern concepts and software, " Experimental Education, Vol.62, No.2, 1994, pp.127–142.

[6] T. K. Landauer, D. Laham, and P. Foltz, "Automated scoring and annotation of essays with the intelligent essay assessor," Automated Essay Scoring: A Crossdisciplinary Perspective, 2003, pp.87–112.

[7] S. Elliot, "IntelliMetric: From Here to Validity," Automated Essay Scoring: A Crossdisciplinary Perspective, 2003. pp.71–86.

[8] V. S. Kumar and D. Boulanger, "Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined ?," Artificial Intelligence in Education, Vol.31, 2021, pp.538–584.

[9] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," Artificial Intelligence Review, Vol.55, 2022, pp.2495–2527.

[10] A. Hiramatsu and T. Yamasaki, "Comparison of Similarity Calculation Methods Using Graph Representation for Scoring Summary Documents for Essay Learning," IEEJ Annual Conference on Electronics Information and Systems, GS1-3, 2021, pp.898–903.

[11] K. Takeuchi, Y. Matsumoto, “OCR Error Correction Using Stochastic Morphological Analyzer with Probabilistic Word Model,” IPSJ SIG Technical Report, 1997–NL–121, 1997, pp.17–24.

[12] F. Sato, M. Kitsuregawa, “Improvement of OCR recognition rate in post-processing by combining OCR character probability and pre-trained BERT MASK candidate,” IPSJ SIG Technical Report, Vol.2020–ACC–13, No.3, 2020, pp.1–5.

[13] S. Zhang, H. Huang, J. Liu, and H. Li, “Spelling Error Correction with Soft-Masked BERT,” Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 2020, pp.882–890.

[14] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, “Masked Language Model Scoring,” Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 2020, pp. 2699–2712.

[15] M. Kaneko, M. Mita, S. Kiyono, J. Suzuki, and K. Inui, “Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction,” Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 2020, pp. 4248–4254.

[16] Japanese Technical Committee for Optical Character Recognition, ETL-9B Character Database, 1973–1984.