Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ensemble Method for OCR Error Handling in Early-Modern Japanese Texts

Hiyori Kanetaka

kanetaka-hiyori0203@lics.nara-wu.ac.jp

Nara Women's University

Yu Ishikawa

Kyoto Women's University

Miho Chiyonobu

Shiga University

Yuki Takemoto

Kyoto Women's University

Masami Takata

Nara Women's University

Research Article

Keywords: Error Correction, Early, Modern Japanese Books, Character Recognition, Bi--GRU

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11275171/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Ensemble Method for OCR Error Handling in Early–Modern Japanese Texts

Hiyori Kanetaka[1*], Yu Ishikawa[2], Miho Chiyonobu[3], Yuki Takemoto[2], Masami Takata[4*]

[1*]Graduate School of Humanities and Sciences, Nara Women's University, Kitauoya–Nishi–machi, Nara, 6308506, Nara, Japan.
[2]Faculty of Data Science, Kyoto Women's University, Higashiyama, Kyoto, 6058501, Kyoto, Japan.
[3]Data Science and AI Innovation Research Promotion Center, Shiga University, Baba 1, Hikone, 5228522, Shiga, Japan.
[4*]Research Group of Information and Communication Technology for Life, Nara Women's University, Kitauoya–Nishi–machi, Nara, 6308506, Nara, Japan.

*Corresponding author(s). E-mail(s): kanetaka-hiyori0203@lics.nara-wu.ac.jp; takata@ics.nara-wu.ac.jp;
Contributing authors: ishikaway@kyoto-wu.ac.jp; chiyonobu-miho@biwako.shiga-u.ac.jp; takemotoy@kyoto-wu.ac.jp;

Abstract

Recently, the digitization of books has been actively promoted to improve accessibility. However, early modern Japanese books published from the Meiji to early Showa era (1868–1935) exhibit numerous characteristics that differ from contemporary publications. Current research continues to advance specialized character–recognition systems for these books. Nevertheless, recognition errors remain prevalent. In this study, we introduce a method for textual error detection and correction in Optical Character Recognition (OCR)–processed early–modern Japanese texts. The proposed approach employs an ensemble framework that integrates a Bidirectional Gated Recurrent Unit (Bi–GRU) model and a dictionary for error detection, alongside BERT for error correction. Evaluation experiments demonstrate that our method achieves an F–measure of **74.9%**, outperforms baseline methods, and effectively identifies erroneous characters with high accuracy while minimizing false detections.

Keywords: Error Correction, Early–Modern Japanese Books, Character Recognition, Bi–GRU

1 Introduction

Paper–based media are inherently susceptible to physical degradation, and accessing or searching their contents is time–consuming. Accordingly, digital conversion has become increasingly necessary. The primary method of conversion is text extraction through Optical Character Recognition (OCR) software, which converts printed or handwritten text—captured by applications or dedicated devices—into digital text data by identifying textual components within image data. Existing OCR systems have primarily been developed for standardized contemporary characters; consequently, they exhibit reduced accuracy when applied to text from pre–standardization eras.

One particularly challenging category for existing OCR systems comprises the characters found in early–modern Japanese publications, produced from the Meiji era to the early Showa era (1868–1935). Unlike contemporary books, these publications lack uniform standards, with fonts varying across publishers and publication years, which substantially reduces recognition accuracy in existing OCR systems [1]. To address this issue, specialized OCR methods employing Convolutional Neural Networks (CNNs) [2] have been proposed [3, 4]. For example, [3] reports the development of an OCR program capable of extracting text data from digitized copies of books preserved in the National Diet Library, achieving recognition accuracy exceeding 90% [5]. Furthermore, [4] employed training data combining modern printed characters with manually collected samples from early–modern Japanese book yielding approximately 97% accuracy. Given the extensive quantity of early–modern Japanese books, manual error detection remains highly difficult. Recognition accuracy can be enhanced through diversification and expansion of training data. Accordingly, research has explored character image modernization and augmentation of character variety [6, 7]. However, the recognition accuracies achieved through these approaches remains limited.

Conventional approaches to recognition error correction include [8] and [9]. [8] identifies the tendency of erroneous sentences to appear unnatural and proposes a three–stage framework comprising erroneous sentence detection, recognition error detection, and recognition error correction. In contrast, [9] adopts a two–stage framework of error detection and correction. Both works report performance evaluations demonstrating the effectiveness of their error correction methods. However, while real–world data contain a natural mix of correct and erroneous sentences, the evaluation dataset in [8] consisted exclusively of erroneous sentences. Although [9] employed mixed data, the dataset contained a relatively large proportion of errors, which differs from the natural distribution of errors expected in real–world data. Therefore, further evaluation under more realistic error distributions is needed to assess the practical performance of these approaches.

In this study, we target texts containing both correct and erroneous sentences. Our proposed method employs a two–stage architecture. For error detection, to mitigate

false positives in correct sentences, we introduce an ensemble approach that integrates dictionary–based deterministic word matching with context–based probabilistic detection [9] using a Bidirectional Gated Recurrent Unit (Bi–GRU) [10]. For error correction, the detected erroneous characters are replaced with a [MASK] token, and the masked tokens are predicted using Bidirectional Encoder Representations from the Transformers (BERT) Masked Language Modeling (MLM) task [11]. The predicted Chinese characters are then selected as correction candidates through a comprehensive evaluation of their string similarity with the original erroneous characters.

2 Character Recognition for early–modern Japanese books

One representative OCR system for early–modern Japanese books is the NDLOCR, developed by the National Diet Library [3]. The system has been fine–tuned on approximately 2.62 million images to adapt to the layout and character features of modern books, employing a CNN integrated with Connectionist Temporal Classification (CTC) [12]. In evaluation experiments using scanned books and magazines published between 1870 and 1990, NDLOCR achieved recognition accuracy exceeding 90%.

Numerous methods have been proposed to enhance character–level recognition accuracy and mitigate the scarcity of training data. Yasunami et al. [4] trained a CNN using modern fonts combined with a limited set of images from early–modern Japanese books, thereby addressing the blurring and degradation inherent in letterpress printing. This method was evaluated on six categories of character images extracted from early–modern Japanese books published across different eras and publishers, achieving recognition accuracies ranging from 95.98% to 97.54%. To address the shortage of training data for specific character classes, Koiso et al. [13, 14]introduced a data–augmentation method employing CycleGAN [15] to generate character images in the style of early–modern Japanese books. This method was evaluated using real character images excluded from the training and validation datasets. The results demonstrated that even with only a single real character image, the inclusion of generated images yielded a maximum recognition accuracy of 94.90%. The method achieved 87.99% maximum recognition accuracy even without using any real character images. In subsequent work, Koiso et al. [16, 17] further proposed replacing the recognizer with a Vision Transformer (ViT) [18]. Character recognition experiments were conducted on 6,524–character classes, encompassing the combined JIS Level 1 and Level 2 –character sets. Their proposed ViT–based recognizer achieved a maximum recognition accuracy of 94.65% with diffusion–generated images and 93.19% with CycleGAN–generated images, thereby substantiating its effectiveness.

Thus, character recognition technology for early modern Japanese books has made significant advances in recent years. However, even when the recognition accuracy exceeds 90%, several characters per hundred remain misrecognized. Therefore, correcting these errors is essential to make digitized materials suitable for practical use as searchable text data and corpora.

3 Previous research

[8] and [19] employed a BERT–based contextual analysis framework to detect and correct recognition errors in early–modern Japanese book texts. Their approach comprised three distinct stages: erroneous sentence detection, recognition error detection, and recognition error correction. For erroneous sentence detection, the authors utilized a BERT–based classifier. The pretrained BERT model was independently developed using early–modern Japanese book data. In addition, erroneous sentences were artificially generated by inserting recognition errors into the corpus. The model was then fine–tuned with labeled pairs of correct and erroneous sentences to construct a classifier capable of binary classification.

For erroneous sentence detection, the authors utilized a BERT–based classifier. The pretrained BERT model was independently developed using early–modern Japanese book data. In addition, erroneous sentences were artificially generated by inserting recognition errors into the corpus. The model was then fine–tuned with labeled pairs of correct and erroneous sentences to construct a classifier capable of binary classification.

Recognition–error detection was performed via dictionary–based word matching. In addition to *jukugo* extracted from the pre–training dataset using SentencePiece, the dictionary incorporated *jukugo* formed by concatenating consecutive Chinese characters. In Japanese, *jukugo* denotes a compound word composed of two or more Chinese characters that collectively convey a unified meaning. For sentences identified as erroneous during the erroneous–sentence–detection stage, word segmentation was performed and cross–referenced against the constructed dictionary. Words absent from the dictionary were classified as recognition errors.

For recognition error correction, correction candidates were generated by integrating filtering mechanisms with the Mask Language Modeling (MLM) task of the pretrained BERT model employed in the detection phase. The procedure first extracts the initial and terminal Chinese characters of *jukugo* containing recognition errors. The position of the recognition error is then replaced with a [MASK] token. Subsequently, word candidates for the [MASK] token are predicted in descending order of their occurrence probabilities using the MLM task. From these predictions, only those beginning with the initial Chinese or ending with the terminal Chinese character of the original recognition error are retained. The top n words obtained through this process are designated as correction candidates.

When creating the training data for erroneous sentence detection in [8], erroneous sentences were generated by randomly replacing Chinese characters. In actual character recognition, however, misrecognitions typically occur between Chinese characters of similar shapes. Consequently, recognition error detection in real–world environments is more challenging than when using randomly substituted recognition errors. Both correct and erroneous sentences were included for the test data; however, no justification was provided for the ratio between them.

4 Proposed Method

This study proposes a method for detecting and correcting character recognition errors in early–modern Japanese books. The approach integrates a Bi–GRU model with

a dictionary–based method in an ensemble framework for error detection. The Bi–GRU model facilitates probabilistic judgments by leveraging contextual information and can reasonably accommodate unknown words. However, it exhibits sensitivity to low–frequency words in the training data, thereby increasing false detections. In contrast, the dictionary–based method performs deterministic matching. While reliably recognizing words registered in the dictionary, this method tends to misidentify words absent in the dictionary. Because these two approaches exhibit different sources of false detections, our proposed method adopts an ensemble technique in which each compensates for the errors of the other, thereby reducing false detections in a complementary manner. Furthermore, we employed evaluation data specifically constructed to reflect actual character recognition error rates, enabling validation of the effectiveness of our proposed method under real–world operational conditions. Our proposed method comprises two stages: error detection and error correction. In the OCR of early–modern Japanese books, characters with limited training data or those with similar forms are frequently misrecognized, a tendency particularly pronounced in Chinese characters. Accordingly, this study focuses on detecting and correcting recognition errors in Chinese characters.

Fig.1 illustrates the overview of the proposed method. In the error detection stage, the system identifies erroneous locations in the input text. In the subsequent error correction stage, the system predicts the correct word for each detected location and revises the input text accordingly.

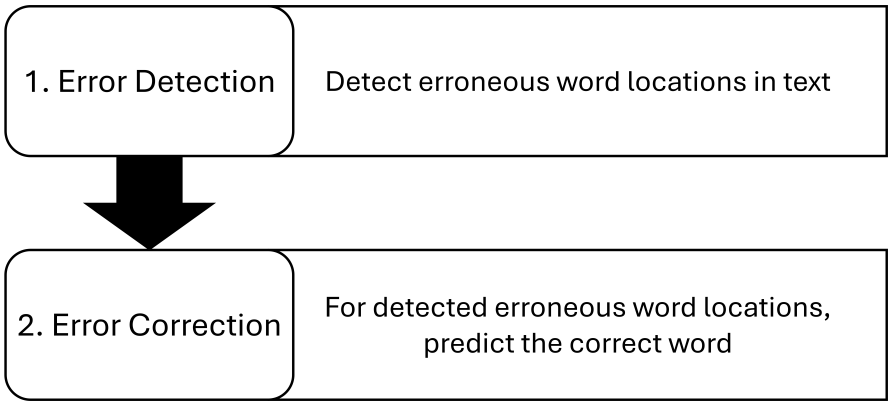


Fig. 1 Flow of our proposed method

4.1 Generation of Erroneous Sentence Data

Fig. 2 illustrates the procedure for generating erroneous sentence data. First, the correct sentence was tokenized using SentencePiece (2–A). From the resulting token sequence, *jukugo* comprising two or more Chinese characters were extracted as replacement targets (Fig. 2–B). Subsequently, one character within the extracted *jukugo* was substituted with a different Chinese character to induce a recognition error (Fig. 2–C). At this stage, to approximate actual OCR misrecognitions, visually similar Chinese

characters identified using ResNet–18, a CNN, were employed as replacement candidates. This approach yields erroneous data that closely replicate real–world OCR recognition errors. Examples of OCR misrecognition errors caused by visually similar characters are shown in Fig. 3. After re–embedding the generated erroneous character into the original sentence (Fig. 2–D), correctness labels were assigned to each token (Fig. 2–E). Specifically, an error label of 1 was assigned to the entire token of any *jukugo* containing an error, whereas a label of 0 was assigned to all other correct tokens.

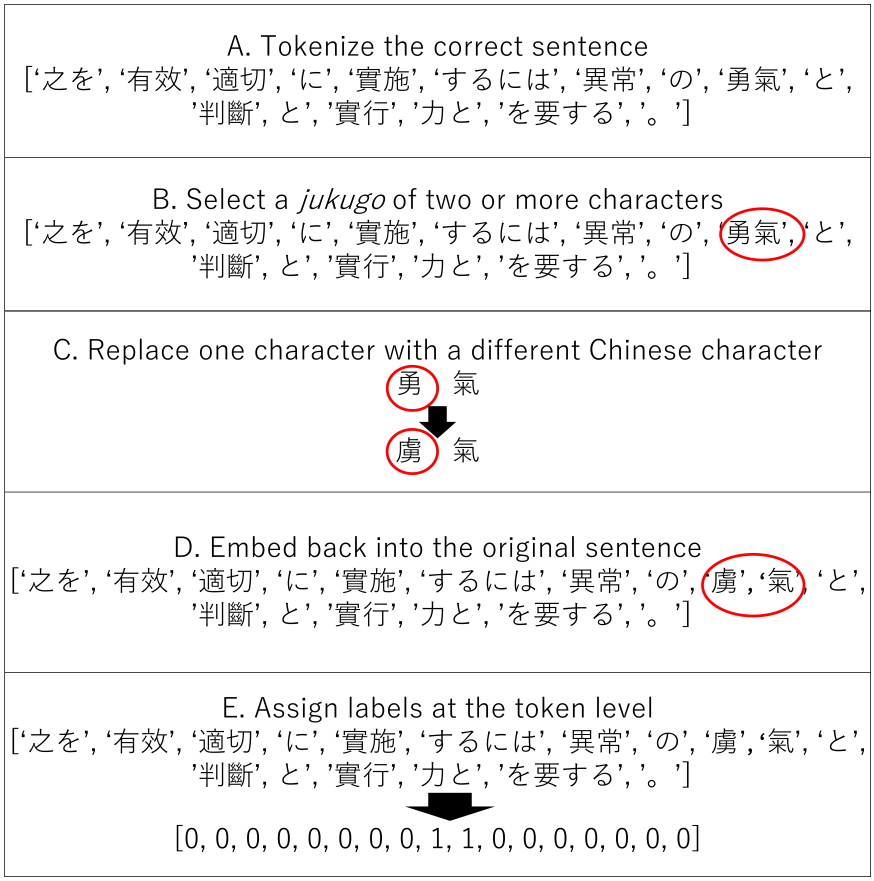


Fig. 2 Procedure for generating erroneous sentences

Pre-processed image data	Correct character	Recognized character
	面	四
	問	凹
	年	字
	羨	美
	寧	發

Fig. 3 Examples of visually similar OCR misrecognitions

4.2 Recognition Error Detection Method

In the recognition error detection, words containing OCR misrecognitions were detected. Target recognition errors for detection are restricted to Chinese substitution errors, wherein a Chinese character within a sentence is erroneously recognized as a different Chinese character. In this study, we propose an ensemble approach that integrates a probabilistic detection method based on Bi–GRU with a deterministic detection method utilizing a dictionary.

The dictionary–based detection method was implemented following the approach described in [8].

The Bi–GRU detection model determines whether each token in a sentence constitutes a recognition error. The input text was tokenized using SentencePiece, with special tokens [CLS] and [SEP] appended to the beginning and end of the sentence, respectively. The resulting token sequence was converted into vector representations and processed by the Bi–GRU. Utilizing the Bi–GRU enables feature extraction that incorporates contextual information from both forward and backward directions for each token. The Bi–GRU output was subsequently passed to a fully connected layer to perform binary classification, thereby determining whether each token was correct or erroneous. For model training, labeled data comprising shuffled correct and erroneous sentences was employed. The generation of erroneous sentences followed the procedure outlined in Section 4.1.

Because dictionary–based detection methods classify all unknown words absent from the dictionary as errors, they tend to produce excessive false positives. Conversely, Bi–GRU–based detection models infer from contextual information, thereby exhibiting robustness against unknown words. However, their detection accuracy becomes unstable for special words or proper nouns with low appearance frequencies in the training data.

Accordingly, this study adopts an ensemble method that integrates a Bi–GRU–based detection model with a dictionary–based detection method. Specifically, dictionary–based detection was first applied to extract candidate errors from sentences. Subsequently, the Bi–GRU model computed error probabilities for these candidates. Finally, candidates exceeding a predefined threshold were identified as recognition errors.

4.3 Recognition Error Correction Method

In recognition error correction, candidates for the correct word are proposed at the positions of the recognition errors. The ranking of these correction candidates is determined by an integrated score that combines the MLM score from the pretrained BERT model with the string similarity to the original recognition error. Furthermore, restricting the candidates to those of identical string length as the recognition error suppresses replacements with inappropriate vocabulary. The pretrained BERT model used for generating correction candidates was constructed following the procedure described in [8].

Correction was performed on the detected recognition errors using the following procedure, and the top n words obtained were designated as correction candidates.

1. The location of the recognition error was replaced with a [MASK] token.
2. The occurrence probability S_{MLM} of the tokens for the [MASK] is calculated using the MLM task of the pre–trained BERT model.
3. The string similarity between the original error token and each candidate token was evaluated using Difflib, a standard Python library

$$S_{SIM} = \frac{2 \times M}{T}, \tag{1}$$

   where T denotes the total number of characters across the two tokens, and M the number of matching characters between them. The similarity S_{sim} is then computed according to Equation1:
4. The comprehensive score is calculated using Equation2 below to rank each candidate:

$$S_{total} = w \cdot S_{MLM} + (1 - w) \cdot S_{sim}. \tag{2}$$

5. From these candidates, tokens whose string length matched that of the detected recognition error were retained as the final correction candidates.

5 Experiment

In this section, to assess the effectiveness of the proposed method, we conducted experiments on recognition error detection and correction using early–modern Japanese books. Section 5.1 details the dataset, Section5.2 presents the results of recognition error detection, and Section 5.3 presents the results of recognition error correction.

5.1 Dataset

The dataset employed in the experiments comprised digital texts of early–modern Japanese books generated by NDLOCR, which are publicly available in the National Diet Library's Next Digital Library [20]. Each text file is formatted in such that one line corresponds to a single page. Because the original data contained unreadable characters and extraneous symbols, preprocessing was performed prior to use. Specifically, extraneous symbols and unnecessary character strings were removed, and sentences containing page numbers or unreadable characters were excluded from analysis. After preprocessing, 282,893 sentences were obtained.

For the experiments, 280,000 sentences were selected from the preprocessed data. Among these, 80% (224,000 sentences) were used as training data, 10% (28,000 sentences) as validation data, and the remaining 10% (28,000 sentences) as evaluation data. To train BERT on the MLM task, 224,000 training sentences were used without modifications. For the Bi–GRU–based detection model, recognition errors were artificially generated from the BERT training data using the method described in Section 4.1. The resulting training dataset comprised 223,999 correct sentences and 223,999 erroneous sentences, yielding a total of 447,998 sentences. Additionally, 119,883 *jukugo* words extracted from the BERT training data were stored in the dictionary.

For the evaluation data, the ratio of correct to erroneous sentences was determined based on the recognition error probability reported in [4], which estimated that the recognition error rate for early modern Japanese books is approximately 3%. In the dataset used in this study, the average sentence length was approximately 54 characters, and the proportion of Chinese characters was approximately 42%. Each sentence contained an average of approximately 22.7 Chinese characters. Assuming a recognition error probability of 3%, the probability that a sentence contained no recognition error was estimated to be approximately 50.1%, implying that nearly half of the sentences contained at least one recognition error. Based on this estimation, an approximately 1:1 ratio of correct to erroneous sentences was adopted for the sentence–level evaluation. Consequently, the evaluation dataset consisted of 53,938 shuffled sentences, including 25,938 erroneous sentences generated using the method described in Section 4.1 and 28,000 correct sentences.

5.2 Results and Discussion

For recognition error detection, the effectiveness of the proposed method, which combines the Bi–GRU–based detection model with a dictionary–based method, was evaluated by against both the Bi–GRU–only and dictionary–only methods. Based on preliminary experimental results, this experiment sets the threshold for the standalone Bi–GRU model to $\theta = 0.75$ and the proposed method's threshold to $\theta = 0.55$. The evaluation was conducted at the token level using four metrics: Accuracy, Precision, Recall, and F–measure. Accuracy denotes the proportion of tokens correctly classified as either correct or erroneous. Precision denotes the proportion of detected recognition errors that are true errors. Recall denotes the proportion of actual recognition errors

that are successfully detected. The F–measure is defined as the harmonic mean of Precision and Recall. True Positives (TP), True Negatives (TN), False Positives(FP), and False Negatives (FN) were defined as follows:

$$
\begin{aligned}
TP &: \text{Number of error tokens correctly classified as error tokens} \\
TN &: \text{Number of tokens correctly classified as correct tokens} \\
FP &: \text{Number of correct tokens misclassified as error tokens} \\
FN &: \text{Number of error tokens misclassified as correct tokens}
\end{aligned}
$$

$$\text{Accuracy} : \frac{TP + TN}{TP + FP + TN + FN}, \tag{3}$$

$$\text{Precision} : \frac{TP}{TP + FP}, \tag{4}$$

$$\text{Recall} : \frac{TP}{TP + FN}, \tag{5}$$

$$\text{F–measure} : \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{6}$$

Table 1 Comparison of recognition error detection results [%].

Method	Accuracy	Precision	Recall	F–measure
Dictionary–only method	88.4	40.2	86.7	55.0
Bi–GRU–only method	95.2	70.7	70.8	70.8
Ensemble method	96.0	72.5	77.4	74.9

The experimental results are presented in Table 1. Table 1 shows that the proposed method achieved superior performance compared to the other methods, with an accuracy, precision, and F–measure of 96.0%, 77.4%, and 74.9%, respectively. However, recall was limited to 72.5%, lagging behind some comparison methods.

To evaluate the detection accuracy at the sentence level in greater detail, the output results for erroneous sentences were classified into four categories:

- Perfect Match: All errors in the sentence were detected exactly, without any excess or omissions.
- Over–detection: All errors in the sentence were successfully detected, but some correct tokens were incorrectly flagged as errors.
- Incorrect: Errors are only partially detected, or correct tokens were misclassified as errors.
- Missed: No errors detected.

The classification results are listed in Table 2. Table 2 indicates that the proposed method achieved a 51.6% accuracy rate for exact matches, demonstrating its ability

Table 2 Sentence–level classification results of error detection [%].

Method	Perfect Match	Over-detection	Incorrect	Missed
Dictionary–only method	48.9	30.7	11.6	9.0
Bi–GRU–only method	47.2	8.6	27.3	16.9
Ensemble method	51.6	9.0	19.9	19.6

to identify misspelled words without omission or over-detection compared with the comparison methods. However, the proportion of missed detections increased to 19.6% compared with the other approaches.

Table 3 The proportion of over–detections in correct sentences [%].

Method	Over–detection in correct sentences
Dictionary–only method	27.7
Bi–GRU–only method	23.3
Ensemble method	13.3

Table 3 presents the proportion of over–detections in correct sentences, in which one or more locations were incorrectly detected as errors, despite the absence of any actual errors. The proposed method achieved the lowest excess detection rate (13.3%) for the correct sentences.

5.3 Recognition error correction

For recognition error correction, the error locations identified by each of the three detection methods compared in the previous section were targeted for correction, and the correction method described in Section 4.3 was applied. This evaluates the impact of detection accuracy on the subsequent correction performance. In this study, recognition error correction involved presenting candidates for the correct word at the locations of the recognition errors. In the calculation of the S_{total}, the weight parameter w was set to 0.7 for all experiments. The top 5–word candidates for each recognition error location were outputted as correction candidates. If the correct word was included among the candidates, the correction was considered successful. In addition, the proportion of incorrectly detected correct sentences recovered without corruption is presented as the recovery rate.

Table 4 Results of recognition error correction [%].

Method	Correction success rate	Recovery rate
Dictionary–only method	–	–
Bi–GRU–only method	57.5	52.4
Ensemble method	64.7	48.4

Table 4 presents the recognition error correction accuracy when the sentences processed using each detection method were used as inputs. The dictionary–based method generated numerous over–detections, resulting in multiple [MASK] tokens being assigned to a single sentence. Consequently, the number of correction candidates for multiple [MASK] tokens increased drastically, leading to a combinatorial explosion; therefore, this method was excluded from the current evaluation. As shown in table 4, the proposed method successfully corrected 64.7% of the input sentences, demonstrating a 7.2–point improvement over the performance of the Bi–GRU model. However, in terms of correct sentence recovery rate, the proposed method exhibited a 4% lower performance.

6 Discussion

This section presents the analysis and discussion of the results. Section 6.1 discusses the effectiveness of the proposed method, and Section 6.2 addresses the remaining challenges.

6.1 Effectiveness of the proposed method

In this section, we discuss the effectiveness and advantages of the proposed method. For error detection, the method achieved[A34.1] higher accuracy and precision than the comparison methods. This improvement is attributed to the ensemble strategy, which integrates the contextual judgment of the Bi–GRU model with the deterministic matching of the dictionary–based method. The Bi–GRU model is sensitive to low–frequency words and proper nouns, often resulting in an increased number of false detections, whereas dictionary–based methods tend to misclassify unknown words that are not registered in the dictionary. By combining these two approaches, false detections generated by one method can be filtered by the other, thereby reducing the overall number of false detections. This complementary error–suppression mechanism enhances the reliability of error detection and contributes to the higher accuracy and precision of the proposed method.

Method	Output Excerpt	Result
Bi-GRU-only method	...いはゆる地方**玫權**化し去ることは事實である。	Over-detection (FP) on *jukugo* “地方".
Ensemble method	...いはゆる地方**玫權**化し去ることは事實である。	FP suppressed, while true error is kept.

Fig. 4 Detection results of using Bi–GRU and our proposed method

Examples of the detection results are presented in Fig 4 and 5. Fig 4 compares the detection results of the Bi–GRU model with those of the proposed method. In the figure, recognition errors are highlighted in bold, and detected locations are underlined. The standalone Bi–GRU approach identifies recognition errors and[A35.1] incorrectly identifies correctly recognized words as erroneous. In contrast, the proposed method successfully detected error locations while suppressing false positives,

Method	Output Excerpt	Result
Dictionary-only method	…畫圖に就ひて考證するの外、他策なければなり。	Over-detection (FP) on *jukugo* "他策".
Ensemble method	…畫圖に就ひて考證するの外、他策なければなり。	FP suppressed

Fig. 5 Detection results of the dictionary–based and proposed method

as the dictionary–based component rejected correctly recognized words. Fig 5 compares the detection results of the dictionary–based method with those of the proposed method. Although dictionary–based detection misclassifies correctly written *jukugo* as errors, the proposed method effectively reduces false detections because the Bi–GRU model filters out such misclassifications. These results demonstrate that the proposed method achieves reduced false detections by the complementary use of both contextual and dictionary matching information compared with single–method approaches.

Table 5 Correction success rate by detection type [%].

Detection type	Correction success rate
Perfect match	91.4
Over–detection	15.2
Incorrect	18.4

A key factor contributing to the improved error correction accuracy is that more target error locations were detected as exact matches during the error–detection stage. Table 5 lists the correction success rates by detection type for the proposed method. Exact matching enables the precise identification of error locations without omission or over–detection, clearly defines correction targets, and reduces unnecessary candidate generation. Consequently, the correction success rate exceeded 90% when the exact matches were achieved. In contrast, over–detection and incorrect cases involved detection ranges that encompassed partial or multiple components of *jukugo* or included correctly recognized words. In such cases, the original string information is partially or fully masked, disrupting the contextual reasoning capabilities of the correction model. This often leads to incorrect correction candidates that fail to maintain contextual consistency, resulting in an accuracy of less than 20%. The proposed method achieved a higher proportion of exact matches than the comparison methods, thereby enabling more appropriate selection of correction candidates and contributing to improved correction accuracy.

6.2 Issues

Although the proposed method improved the error detection performance and consequently contributed to a higher error correction accuracy, several limitations remain. The proposed method exhibited reduced recall during error detection, in some cases,

both error correction and restoration of the correct text failed. The following subsections discuss the causes of these issues and outline possible directions for future research. First, we discuss the decrease in the recall. The proposed method suppresses overdetection by identifying only those locations identified as errors by both the Bi–GRU model and the dictionary–based method. However, this strategy excludes errors detectable by only one of the two methods, resulting in an increased number of missed detections, and consequently, lower recall. This reflects the tradeoff between improving precision and maintaining recall. As reducing missed detections remains an important challenge, future studies should explore more flexible detection criteria, such as dynamically adjusting dictionary–based decisions according to the Bi–GRU model confidence score, to maintain recall while suppressing false detections. Next, we discuss the causes of failure in error correction and restoration of correct text. Analysis of unsuccessful cases revealed two major factors.

The first factor was the presence of words that were neither registered in the dictionary nor frequently observed in training data. Although the proposed method substantially reduced false detections, errors still occurred for such low–frequency words because both the Bi–GRU model and dictionary–based method lacked sufficient knowledge.

The second factor was the mismatch between the actual error location and the detected range. As shown in the previous section, such mismatches directly reduce correction success because the correction model receives an incorrect correction target. Furthermore, although the proposed method successfully reduced the over–detection of correct text, it resulted in a lower recovery rate for the correct text. Because the same correction model was applied to both the proposed and comparison methods, this difference is attributable not to the correction model itself but to the characteristics of the input passed to the correction stage. The proposed method effectively suppresses unnecessary detections of ordinary correct text, leaving primarily difficult cases, such as contextually ambiguous sentences and low–frequency words, including proper nouns, as correction targets. Because these cases are inherently more challenging for the correction model to restore correctly, the recovery rate of the correct text decreases. In addition to expanding the training data and dictionary, the introduction of an error detection method capable of identifying error locations more precisely is expected to further improve both the detection and correction performances.

7 Conclusion

In this study, we propose a recognition error detection method based on an ensemble of probabilistic detection methods using a Bi–GRU model, a deterministic detection method using a dictionary, and a recognition error correction method using BERT. Existing OCR systems do not achieve sufficient accuracy for early modern Japanese books. Although OCR methods specialized for early modern Japanese books have been proposed, their recognition accuracy remains imperfect. Therefore, recognition error detection and correction are required for text data after character recognition. The proposed method comprises two stages: recognition error detection and correction. The experimental results demonstrate that the proposed method achieves a higher

detection performance than single–model approaches and contributes to improved accuracy in the subsequent recognition error correction stage.

Acknowledgment

This study was supported in part by JSPS KAKENHI, Grant Number JP26K21587, and the Initiative for Realizing Diversity in the Research Environment. We would like to thank Editage (www.editage.jp) for English language editing.

Declarations

Funding

JSPS KAKENHI, Grant Number JP26K21587 and the Initiative for Realizing Diversity in the Research Environment.

Ethics approval

Not applicable.

Availability of data and materials

The data used in the paper will be available upon request.

References

[1] Ashida N, Takata, M Sasaki A, Kamo H, Nide N, Joe K (2007) Construction of a Paper Classification System Using SVM, Proceedings of The 2007 International Conference on Parallel and Distributed Processing Technologies and Applications 2:759–765.

[2] LeCun Y, Bottou L, Bengio Y, Haffiner P (1998) Gradient–based learning applied to document recognition. Proceedings of the Institute of Electrical and Electronics Engineers 86(11):2278–2324. https://doi.org/10.1109/5.726791

[3] NDLOCR, https://github.com/ndl-lab/ndlocr_cli/. Accessed 24 Aug 2026

[4] Yasunami S, Koiso N, Takemoto Y, Ishikawa Y, Takata M, Joe K (2019) Applying CNNs to Early–Modern Printed Japanese Character Recognition. Proceedings of The 2019 International Conference on Parallel and Distributed Processing Technologies and Applications. pp 189–195

[5] OCR text conversion of digitized materials in FY2021, https://lab.ndl.go.jp/data_set/ocr_en/r3_text/

[6] Kosaka K, Awazu T, Ishikawa Y, Takata M, Joe, K(2015) An Effective and Interactive Training Data Collection Method for Early–Modern Japanese Printed

Character Recognition. Proceedings of The 2015 International Conference on Parallel and Distributed Processing Techniques and Applications 2:276–282.

[7] Fujisaki N, Ishikawa Y, Takata M, Joe K (2021) Crawling Low Appearance Frequency Characters Images for Early–Modern Japanese Printed Character Recognition. Advances in Parallel & Distributed Processing, and Applications. pp 683–695.

[8] Kanetaka H, Chiyonobu M, Takemoto Y, Ishikawa Y, Takata M (2025) Correction of textual errors in early–modern Japanese books. J Supercomput 81:1592. https://doi.org/10.1007/s11227-025-08076-y

[9] Zhang S, Huang H, Liu J, Li H (2020) Spelling Error Correction with Soft-Masked BERT. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp 882–890. https://doi.org/10.18653/v1/2020.acl-main.82

[10] Cho, K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. pp 1724–1734. https://doi.org/10.3115/v1/D14-1179

[11] Devlin J, Chang M.–W, Lee K, Toutanova K (2019) BERT: Pre–training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1:4171–4186. https://doi.org/10.18653/v1/N19-1423

[12] Graves, A, Fernández S, Gomez F, Schmidhuber J (2006) Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine Learning. pp 369–376. https://doi.org10.1145/1143844.1143891

[13] Koiso N, Takemoto Y, Ishikawa Y, Takata M (2025) Proposed method of acquiring train data for early–modern Japanese printed character recognizers. J Supercomput 81:764. https://doi.org/10.1007/s11227-024-06866-4

[14] Koiso N, Takemoto Y, Ishikawa Y, Takata M (2025) Improved Early–Modern Japanese Printed Character Recognition Rate with Generated Characters. Parallel and Distributed Processing Techniques. CSCE 2024. Communications in Computer and Information Science 2256:167–181. https://doi.org/10.1007/978-3-031-85638-9_13

[15] Zhu J–Y, Park T, Isola P, Efros A A (2017) Unpaired image–to–image translation using cycle–consistent adversarial networks. Proceedings of the IEEE International Conference on Computer Vision. pp 2242–2251. https://doi.org/10.1109/

iccv.2017.244

[16] Koiso N, Chiyonobu M. Takemoto Y, Ishikawa Y, Takata M(2026) Diffusion–Based Data Augmentation for early–modern Japanese Printed Character Recognition. IPSJ Transactions on Mathematical Modeling and Its Applications. (accepted, in japanese)

[17] Koiso N, Chiyonobu M, Takemoto Y, Ishikawa Y, Takata M(2026) Applying Vision Transformer to early–modern Japanese Printed Character Recognition. The 32nd Int'l Conf on Parallel and Distributed Processing Techniques and Applications. (in press)

[18] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2020) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

[19] Kanetaka H, Chiyonobu M, Takemoto Y,Ishikawa Y, Takata M (2026) Correction Method for Text Errors in early–modern Japanese Books Using BERT. Distributed Computing, HPC, Cloud Computing, Quantum Technologies, Embedded Systems, Modeling, and Simulation. CSCE 2025. Communications in Computer and Information Science 2932:417–426. https://doi.org/10.1007/978-3-032-22193-3_32

[20] Next Digital Library, https://lab.ndl.go.jp/dl/. Accessed 24 Aug 2026