Correcting Misuse of Japanese Visually Similar Characters

Youichiro Ogawa and Kazuhide Yamamoto
Nagaoka University of Technology
Nagaoka, Niigata, Japan
{*ogawa, yamamoto*}*@jnlp.org*

***Abstract*—We present a misuse correction method of visually similar Japanese characters, Kanji, based on the language model. While methods for error correction in Japanese learners' writings have been proposed, however the misuse of visually similar Kanji has not been explored yet. We collected pairs or groups of visually similar Kanji and created the similar Kanji set. Then, candidate sentences are generated by replacing the misuse Kanji with similar Kanji extracted from the similar Kanji set, and select the candidate with the highest language model probability. The experimental results suggest that our method showed high performance in many cases of misuse. In addition, using a morphological analyzer, we developed an unknown word filter which excludes candidates that constitute unknown words when generating candidates. We have found that this filter is effective to prevent erroneous corrections.**

***Keywords*-spelling error correction;**

I. Introduction

According to the "Survey Report on Japanese-Language Education Abroad 2015" of the Japan Foundation, approximately 3.65 million people in 137 nations and regions abroad are learning the Japanese language[1]. However, the number of Japanese language educators is one for every 57 learners; therefore, there is a talent shortage in Japanese language education compared to the demand of people interested in learning the Japanese language. For this reason, it is required to support Japanese teachers and learners using an automatic correction system, such as error detection and correction of Japanese writing.

Most research on error correction of Japanese learners' writings has limited a type of errors. In particular, most reports have been limited to error detection and correction of particles [1][2]. However, there are many types of errors such as vocabulary selection, notation and so on, which are indispensable corrections. In addition, there has been an approach that is not limited in terms of error type, which used statistical machine translation for error correction [3]. In this method, a correction model was learned from a large-scale learner parallel corpus including learners' sentences and its corrections. However, the types of errors that could be corrected were limited by the corpus. For example, consider the Kanji "他" and "地" as well as "与" and "写". It was not possible to correct this confusion because the corpus did not include this type of error.

Japanese language has two kinds of characters which are Hiragana and Kanji. Hiragana is a phonogram similar to the English alphabets and the phonemes alone do not express meaning. Hiragana has approximately 50 characters, which are more in number compared to the alphabets. However Kanji is an ideogram where a character has specific meaning. Approximately 3000 characters of Kanji are commonly used in Japan. We need to memorize a large amount of information such as the shape, reading, and meaning of Kanji in order to use these correctly. Therefore, it is difficult for learners to learn Kanji. In addition, learners usually confuse Kanji because of similarity in Kanji shapes. In recent years, technologies such as the handwritten input of letters and reading of handwritten characters using OCR, are developing; therefore, there is a possibility that misuse may be caused by similar shapes. For this reason, we propose a novel method to correct misuse of visually similar Kanji.

[1]https://www.jpf.go.jp/j/project/japanese/survey/result/

II. Related works

Spell check is a traditional and important preprocessing task for natural language processing since spelling errors occur in texts written by non-native speakers. Many research on English spelling error detection and correction has been reported. In English, words are separated by spaces and can be classified as "non-word" if they do not exist in the reference dictionary and "real-word" if it exists. Non-word is detected as a misspelling and corrected to the actual real word. In the conventional method, misspellings are often corrected based on the edit distance. As a more advanced correction method, the use of noisy channel model [4] and context information [5][6] are proposed.

Unlike the English language, Japanese words are not separated by space; therefore, automatic word segmentation is necessary. In Chinese spelling correction, errors have been detected and corrected using a language model [7]. Errors can be detected and corrected by comparing the language model probability of the sentence including the misspelling and the correct sentence. Therefore, We propose a correction method using language model.

III. System Architecture

A. Similar Kanji Set

Pairs or groups of visually similar Kanji are confusing because most of their components are shared and completely identical, with only one or a few components that are actually different. Specifically, similarity in form between Kanji is often caused by these common patterns:

- A stroke that exists in one character but not in the other. Example: "井" and "丼".

978-1-5386-1981-0/17/$31.00

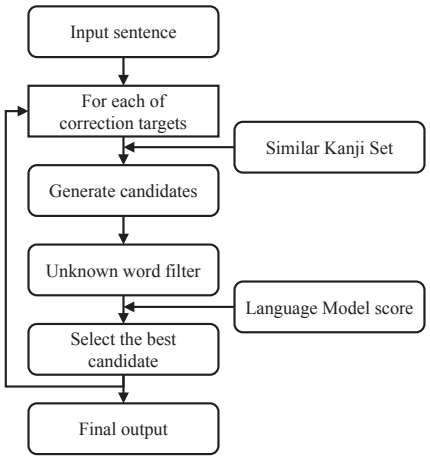

Figure 1. Overview of our system

Table I
STATISTICS OF TWO TEST DATA

Corpus	Learners-error	Pseudo-error
Total sentences	44	10,677
# targets	269	150,277
# errors	47	11,355
Ave. # targets per sentence	6.11	14.1
Ave. # errors per sentence	1.07	1.06

- A stroke that exists in both characters but ends at a slightly different place. Example: “牛” and “午”.
- A different but similar radical. Example: “複” and “復”.
- A different but similar non-radical component. Example: “使” and “便”.

We collected such pairs or groups and created the similar Kanji set. In total, it comprises 608 pairs or groups and has 1,942 characters. As a standard established by the government for the daily usage of Kanji, the commonly used Kanji comprise 2,136 characters[2]. From this amount, this similar Kanji set includes 980 characters.

B. Error Correction

We propose a method to correct for misuse of visually similar Kanji in compositions of Japanese language learners by using a language model. The overview of proposed system is shown in Figure 1.

The Kanji included in the similar Kanji set becomes correction targets. First, by checking characters one by one from the beginning in a sentence, these targets are discovered. If the character is a target, it should belong to one of the pairs or groups in the similar Kanji set. Then, one or more Kanji similar to the target are extracted from the similar Kanji set. And, candidate sentences are generated by replacing the target with each of extracted Kanji. the language model probability is calculated for each candidate sentences. If a candidate sentence with a probability higher than that of the original sentence exists, the candidate sentence with the highest probability is selected. This operation is repeated for all target included in a sentence.

If the target Kanji in a sentence are consecutive, all combinations are listed and each one is replaced with the candidates in order to make them candidate sentences. For example, when “自白 (confession)” is included, the candidate sentence includes “白白”, “自白”, and “自自”.

There are cases where the target Kanji is used as a proper noun. The proper noun is often a word that does not exist in training data, such as a person’s name or a place name. It is difficult to appropriately correct such proper nouns. Therefore, when finding targets in the sentence, the target, classified by morphological analysis as a proper noun, is skipped without any changes being made.

To train our language model, we used the Balanced Corpus of Contemporary Written Japanese(BCCWJ)[3], which contains 104.3 million words. We built 4-gram language model using the KenLM toolkit[4], with modified Kneser-Ney smoothing.

C. Unknown Word Filter

Due to misuse of Kanji, morphological analysis may result in “unknown words” which are words that do not exist in the reference dictionary. Kanji classified to be unknown words cannot be appropriately combined with the surrounding words; therefore, in many cases, it can be judged that they are incorrectly used. For example, in the sentence “約 1 分聞 (間) 打ち続けた (keep on striking for approximately 1 min),”“聞” is classified as an unknown word; however the correct use is ”間,” and therefore, it is judged as misused. Normally, when such an unknown word is included in the sentence, the language model probability should be low. However, the sentence “シドニー五論競泳伐表 (五輪競泳代表) から漏れた (be excluded from the Sydney Olympic swimming team)” contains two misuses and it concatenates with surrounding words. Therefore, “五論競泳伐表” is judged as an unknown word. In this case, the number of words divided by word segmentation is smaller than the number of words in a correct sentence. Furthermore the number of calculations for obtaining the language model probability is reduced; hence the probability is high. In other words, incorrect word segmentation occurs due to misuse of Kanji, which affects the comparison of language model probabilities.

For this reason, as an operation before computing the language model probability, an unknown word filter is inserted. This filter excludes the sentences whose number of characters, classified as unknown words, increases in comparison to the original sentence from the generated candidate sentences. This prevents selecting a wrong candidate sentence.

[2]http://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kijun/naikaku/Kanji/

[3]http://pj.ninjal.ac.jp/corpus_center/bccwj/

[4]https://kheafield.com/code/kenlm/

Table II
PERFORMANCE OF EACH SYSTEM

Corpus	System	R	P	F	#TP	#FN	#FP
Learners-error	Random	0.26	0.071	0.11	12	23	157
	Freq	0.40	0.24	0.30	19	16	60
	LM	0.94	0.88	0.91	44	3	6
	Filter+LM	0.94	0.88	0.91	44	3	6
Pseudo-error	Random	0.339	0.0401	0.0717	3,844	3,794	92,050
	Freq	0.476	0.117	0.188	5,409	2,983	40,740
	LM	0.941	0.856	0.897	10,685	315	1,793
	Filter+LM	0.939	0.883	0.910	10,664	363	1,412

IV. EXPERIMENT

A. Test Corpus

In the experiments, a learners' composition corpus "Natane",[5] and an online Japanese misuse dictionary[6] were used. These are misusage tags attached to compositions collected from Japanese learners. From these, we extracted sentences in which misuse was caused by visually similar Kanji, and collected 44 sentences, including 47 errors, for use in the experiments. However, the case in the learner are considerably few for evaluating the proposed system convenience. In fact, learners make misuse in various Kanji. Therefore, we created a pseudo-error corpus that automatically generated misuse of various Kanji.

Among the misuse and correct sentences, the correct sentences are easily obtained by randomly selecting sentences from the Japanese plain text corpus. Therefore, by replacing the correct Kanji with the similar Kanji, it can be considered as an incorrect sentence, as a learners' composition. We used the Mainichi newspaper corpus (1999-2000) [7]to generate pseudo-error sentences.

The method of generating a pseudo-error sentence is described as follows. First, the Kanji included in the similar Kanji set (the target) is considered. A sentence including this target is extracted from the corpus. Then, one or more Kanji similar to the target are obtained from the similar Kanji set. One of them is randomly chosen, and the target is replaced. If there are several targets in the sentence, replacement should occur at every location. Therefore, there is at least one misuse in the sentence. However, if the target is used as a proper noun or if it becomes a proper noun by replacement, it should not be replaced. When 10 misused sentences are generated from one target, the same operation is performed on the next target. However, when the number of sentences is less than 10, only the same number of sentences is generated. This operation is performed all target Kanji. The generated pseudo-error corpus has 10,677 sentences, including 11,355 errors. The details of these two corpora are summarized in Table I.

B. Evaluation

For evaluation, we used recall (R), precision (P), and F-measure (F). Recall is the rate of valid correction of misused parts, precision is the rate of valid correction for the locations where the system performed correction. and F-measure is the harmonic mean between recall and precision.

[5]https://hinoki-project.org/natane/

[6]http://cblle.tufs.ac.jp/llc/ja_wrong/

[7]urlhttp://www.nichigai.co.jp/sales/mainichi/mainichi-data.html

C. Baseline

For comparison with the proposed method, we set the two baselines described below. Random: This involves the random selection of wrong Kanji from the same pair or group in similar Kanji set. Freq.: This involves the selection the most frequent wrong Kanji from the same pair or group in the similar Kanji set. The frequency of each target Kanji was calculated from the corpus used to train the language model beforehand.

V. RESULTS AND DISCUSSION

We experimented using learner corpus and pseudo-error corpus as test data. The results of the experiment are shown in Table II. True positive (TP) is a location where an accurate correction was performed, false negative (FN) is a location where an misuse was not corrected, and false positive (FP) is a location where an inaccurate correction was made.

First, focusing on baseline score, the score of both corpora results is low. Since the baseline method does not consider the information of surrounding words, it is difficult to select the correct Kanji. On the other hand, the method using the language model scored higher. This indicates that the correct Kanji was selected by utilizing the information of the the preceding and succeeding words.

In the experiment with the learner corpus, using the language model, 44 of 47 mistakes were accurately corrected, with a recall of 0.936. Additionally, 50 locations were corrected, with a precision of 0.880. The F-measure was 0.907, indicating that the system exhibited a high correction accuracy. Even when an unknown word filter was inserted, no change was observed in the score.

In the experiment with the pseudo-error corpus, using the language model, 10,685 of 11,355 mistakes were accurately corrected, with a recall of 0.941. Additionally, 12,478 locations were corrected, with a precision of 0.856. The F-measure was 0.897. Furthermore, combining an unknown word filter considerably reduced the number of FPs, along with an improvement in the precision and the F-measure became 0.910. This was due to the successful removal of candidates, which would cause unknown words to be formed from candidate sentences. Thus, it was shown that the unknown word filter was effective. The recall decreased slightly due to the correctly used parts being

Table III
EXAMPLES OF CORRECTION RESULTS

	Input sentence	Answer	Output
(a)	日本人の矢ったものがここにある。 Here is what the Japanese arrow.	失 lost	失 lost
(b)	「対策の遅れ」を歴史に刻む手にしてしまった。 "Delayed measure" was engraved in history this hand.	年 year	手 hand
(c)	鳥と毛虫と人が共存する樹木。 A tree that birds, caterpillars, and people coexist.	共存 coexist	共有 share
(d)	そして文学をそつぎょうしだいいい日本かいしゃで働いたいです。 And as soon as I graduate from literature I want to work in a good Japanese company.	大学 college	文字 character

occasionally misclassified as unknown words. The pseudo-error corpus had the same performance as the learner corpus, even though there were considerably less instances of errors. Therefore, in many case, the proposed method demonstrated high performance.

The correction examples are shown in Table III. Example (a) corrects the misuse to the correct Kanji, while examples (b), (c), and (d) are erroneous. Example (b) is an example of the system not correcting an error. When calculating the language model probability, it considered not only the frequency information of 4-gram appearing in the training corpus, but also the frequency of 3-gram or less. Since the N-gram around the target in both this example of input and output sentences do not exist in the training corpus, it was simply compared by a word frequency. Example (c) is an example of the system correcting the correct part to an incorrect Kanji. In this case, the change is incorrect, although there is no grammatical error in this example of output sentence. In the method using the language model, only information of several surrounding words was utilized; however this could be improved using a method that considers the meaning of the sentence and the context of the preceding and succeeding sentences. Example (d) is an example of the system correcting a mistake; however, incorrect Kanji is selected. This sentence is a learners' composition and contains many Hiragana. However, the word segmentation of Hiragana may be incorrect. For example, the morphological analysis of "そつぎょう (graduation)" does not judge it as one word, and an incorrect division is performed. This interferes with the accurate correction. Hiragana can also be used instead of Kanji, therefore, ambiguity is high and automatic word segmentation is difficult. In future research, to correct the writings of Japanese language learners and since learners' compositions tend to include many Hiragana, it is necessary to improve the performance of the Hiragana morphological analysis.

VI. CONCLUSION

In this paper, we developed a system to correct the misuse of visually similar Kanji. In the experiment, the learners' composition corpus with few cases and the pseudo-error corpus, including multiple error cases, were used as the test data. Since both corpora could perform correction with high performance, we found that the proposed system can deal with various error cases. However, it is necessary to consider the meaning of the sentence and the context of the preceding and succeeding sentences for cases in which the grammar is correct even when incorrect Kanji is used.

There are also cases in which the morphological analysis error of Hiragana had a decreasing effect on performance. Therefore, in the future improving Hiragana morphological analysis will be necessary for correcting Japanese learners' composition errors.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grants-in-Aid for Scientific Research (B) Grant ID 15H03216.

REFERENCES

[1] K. Imamura, K. Saito, K. Sadamitsu, and H. Nishikawa, "Particle error correction from small error data for japanese learners," *Information and Media Technologies*, vol. 9, no. 4, pp. 834–856, 2014.

[2] K. Imaeda, A. Kawai, Y. Ishikawa, R. Nagata, and d Fumito Masui, "Error detection and correction of case particles in japanese learner's composition(in japanese)," *In Proceedings of the Information Processing Society of Japan SIG*, pp. 39–46, 2003.

[3] T. Mizumoto, M. Komachi, M. Nagata, and Y. Matsumoto, "Mining revision log of language learning sns for automated japanese error correction," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 28, no. 5, pp. 420–432, 2013.

[4] E. Brill and R. C. Moore, "An improved error model for noisy channel spelling correction," *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 286–293, 2000.

[5] P. Fivez, S. Suster, and W. Daelemans, "Unsupervised context-sensitive spelling correction of clinical free-text with word and character n-gram embeddings," *Proceedings of the 16th Workshop on Biomedical Natural Language Processing*, pp. 143–148, 2017.

[6] M. Flor and Y. Futagi, "On using context for automatic correction of non-word misspellings in student essays," *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pp. 105–115, 2012.

[7] J. Yu and Z. Li, "Chinese spelling error detection and correction based on language model, pronunciation, and shape," *Conference on Chinese Language Processing*, pp. 220–223, 2014.