Dataset Augmentation for Grammatical Error Correction Using Markov Chain

Ryoga Nagai and Akira Maeda

***Abstract*— In grammatical error correction (GEC), both corpora before and after proofreading are generally required. In general, these corpora are difficult to obtain. This paper presents the effectiveness of using pseudo-proofread texts generated from a small corpus by the Markov chain model, and using pre-proofread sentences generated using conventional methods such as back translation and rule-based expansion on the pseudo-post-proofread sentences. This study aims at improving the seq2seq model by adding these pseudo sentences to the original corpus.**

***Index Terms*— deep learning, natural language processing, proofreading, seq2seq**

I. Introduction

As global internationalization progresses, Japan is also going through a wave of internationalization. According to the Japan National Tourism Organization (JNTO), the number of foreign tourists visiting Japan in 2019 was 31,882,049 [1]. As only 6,789,658 foreign tourists visited Japan in 2009, it indicates that Japanese society has grown more internationalized in the past 10 years. However, there is a barrier for foreign visitors to Japan: language. According to a survey conducted by the Japan Tourism Agency from 2018 to 2019 on "difficulties encountered while traveling in Japan", "inability to communicate with staff at facilities, etc." was the most common response, accounting for 20.6% of all responses [2]. It can be seen that the language barrier is significant for foreign tourists, and its improvement is important in an internationalized society. Especially with the increase in international communication, the number of foreign students is expected to increase. Sentences that contain errors specific to Japanese or difficulty in reading can be difficult to understand not only for foreigners but also for Japanese people. In these cases, text proofreading has an important role.

Currently, not only beginners of Japanese language, but even professionals such as newspaper journalists are likely to casually use the automatic proofreading functions provided in Microsoft Word or other software. In the past, most of these functions were rule-based by storing editing patterns from a huge corpus, but now some of them are using deep learning [3].

Because deep learning is supervised learning, it requires a large corpus for its training. In particular, for the text proofreading task, it is necessary to prepare both pre- and post-proofread sentences, which are expensive to construct. In this study, we use a Markov chain model to pseudo-extend the post-proofread texts. By using deep learning for classifying sentences, we find and eliminate sentences from the pseudo-extended sentences that are generated by the Markov chain model. Furthermore, we use multiple methods such as rule-based and deep-learning-based back-translation to pseudo-extend the pre-proofread sentences from pseudo and real post-proofread sentences. By doing so, we aim to increase the total amount of the entire corpus and thus improve the accuracy of the proofreading model.

Manuscript received March 23, 2021.

Ryoga Nagai is a master's course student at Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan (e-mail: is0367fs@ed.ritsumei.ac.jp).

Akira Maeda is a Professor of College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan (e-mail: amaeda@is.ritsumei.ac.jp).

. II. Related Work

A. Markov chain model

The Markov property is the property that, the previous past states are irrelevant when predicting the next future state from the current state, and it is independent of the past states [4]. For example, when the states are in the time series $x_1 \dots x_i\{1, 2, \dots, i-1, i\}$, x_{i+1} is determined by $P(x_i|x_{i+1})$. A type of Markov chain model is the *n*th-order Markov chain model. The *n*th-order Markov chain model predicts x_{i+1} using $x_{i-n} \dots x_i$, with *n* being arbitrarily specified. The larger *n* is, the more meaningful the sentence is, but the less originality it has. Markov chain models are sometimes used for corpus expansion. Sentences extended with Markov chain models as the data for machine learning may cause overlearning because the corpus is reused.

B. Deep Learning on Natural Language Processing (NLP)

Overview

A Recurrent Neural Network (RNN) is one of the deep learning models. RNNs have the disadvantage that they cannot refer to long-term word dependencies. Therefore, models that extend the units for storing or forgetting long-term dependencies, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), have been developed.

The deep learning tasks in our study are sentence classification and sentence generation. Sentence classification automatically classifies the input sentences into one of multiple categories. In sentence generation, new data

ISBN: 978-988-14049-2-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

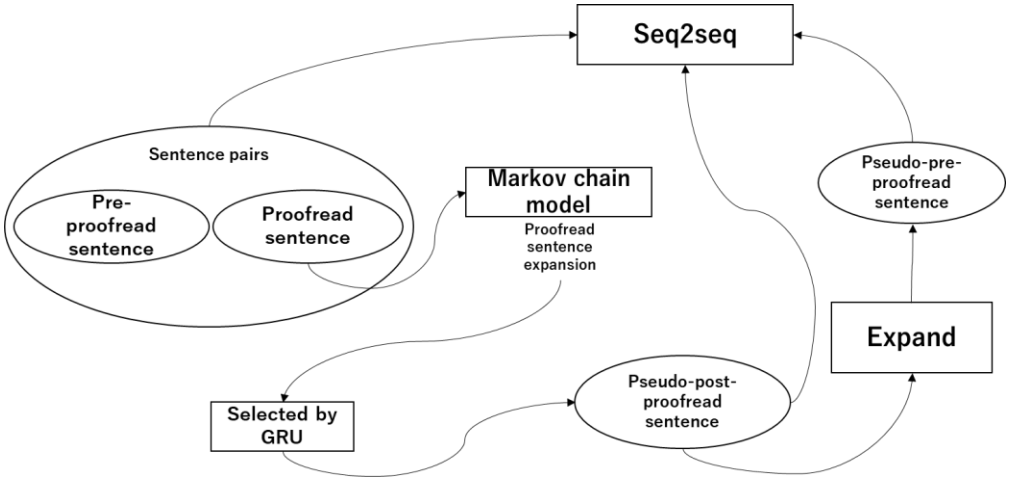

Fig. 1. Overview of the proposed method.

is output from the input time series data. It is widely used in translation and summarization tasks. In this study, we treat sentence proofreading as a translation task that generates correct sentences from incorrect sentences.

Sentence Classification

The sentence classification task using deep learning learns to classify input sentences into one of several categories. There are a wide range of methods, such as Bag of Words, which groups the words in a document together without considering their order. In recent years, document classification using deep learning has become a mainstream method. In this study, we train a bi-directional GRU model on a part of pseudo-proofread texts generated from the Markov chain and actual texts. Using the learned GRU model, we delete "sentences that look like they were generated by a Markov chain" from the pseudo-proofread sentences.

III. Proposed Method

A. Overview

In this study, we expand the size of the corpus by pseudo-generating not only pre-proofread sentences, which is difficult to obtain, but also post-proofread sentences. First, we use an *n*th-order Markov chain model to pseudo-generate the post-proofread sentences. Many of the sentences generated by the Markov chain model contain grammatical errors. To solve this problem, we use a classifier trained on a corpus generated by the Markov chain model and a real corpus to classify the sentences into grammatically reasonable and ungrammatical sentences, and use only the former as the post-proofread sentences. We use multiple methods, such as rule-based and back-translation, to generate pseudo-pre-proofread sentences from the post-proofread corpus including the pseudo-sentences generated by the classifier. We then measure how much the accuracy of the seq2seq proofreading task changes when using the generated corpus and when using only the original corpus. The overall flow of the proposed method is shown in Figure 1. As an accuracy measure, we use BLEU and discuss the results.

B. Expanding Corpus by Markov Chain Model

In the *n*th-order Markov chain model, a large value of *n* results in a loss of originality, and a small value results in a loss of contextual correctness. In this study, we set *n*=2 to emphasize the originality of the generated sentences since we select the sentences by GRU. Examples of sentences generated by the Markov chain model are shown in Table 2

TABLE II
Examples of Markov Chain Output

Number	Example
1	できるだけ早くそれをした。
2	あなたは何になりますか。
3	ボブは会に参加しなかったなかった。
4	彼が試験の結果はまだ生きている。

Table 1 shows that sentences 1 and 2 are semantically correct, while 3 and 4 are incorrect. In this study, we use Markov chains to generate the same number of documents as the corpus.

C. Selecting Expanded Corpus

Markov chain models can generate semantically incorrect sentences, as shown in sentences 3 and 4 in Table 1. If such sentences exist as "post-proofread sentences" in the corpus, they become noise and may reduce the accuracy of the model. Therefore, it is necessary to eliminate such sentences in advance. In this study, we train a classifier using 10,000 simplified Japanese sentences and 10,000 pseudo-proofread sentences generated using 10,000 of the 50,000 sentences in the "SNOW T15: Japanese Simplified Corpus with Core Vocabulary" by Yamamoto et al. [7]. We define ordinary simplified Japanese as "grammatically correct sentences" and pseudo-sentences generated by Markov chains as "grammatically incorrect sentences," and train the classifier as a document classification task. As a result of training, the GRU recorded an accuracy of 68.4%. Using this trained classifier, we classify 40,000 pseudo-post-proofread sentences generated from 40,000 simplified Japanese sentences. Normally, all of them should be classified as artificial sentences, but some of them are classified as real sentences. In this study, the sentences classified as real are added to the 40,000 as a new corpus.

D. Generating Pseudo-pre-proofread Sentences

The extended pseudo-post-proofread sentence does not

ISBN: 978-988-14049-2-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE II
EDITING RULES

Rule	Explanation
Replacement	33% probability of swapping the positions of two words determined by uniform random numbers, 33% probability of doing this operation twice, 34% of doing nothing.
Insertion	For each word, there is a 10% probability that the same word will be inserted one word after it.
Deletion	For each word, there is a 5% probability of deletion.

have corresponding pre-proofread sentence. In this study, we generate pseudo-pre-proofread sentences by three methods: noun and adjective substitution using distributed representation of words, rule-based word substitution and deletion, and back-translation.

E. Generating Pseudo-pre-proofread Sentences by Word2vec

In this study, we use the trained word2vec model available in [8]. We use mecab-ipadic-NEologd [9] to perform morphological analysis on the pseudo-post-proofread sentences, and if the word in a sentence is a noun or an adjective, we replace it with the most similar word.

F. Generating Pseudo-pre-proofread Sentences by Rule-based Method

The rule-based method probabilistically edits the pseudo-post-proofread sentences and generates pseudo-pre-proofread sentences. In this study, the rules are set as shown in Table 2. The probability of each rule is independent of each other, and processing is done on a sentence-by-sentence basis.

G. Generating Pseudo-pre-proofread Sentences by Back Translation

In the sentence proofreading task, it is common to train a neural network to generate post-proofread sentences from pre-proofread sentences. In sentence expansion, the opposite input and output is used to generate a pseudo-pre-proofread sentences. This task is called the back translation. Ogawa et al. [10] showed that data expansion by back translation can improve the accuracy of deep learning models. Gu et al. [11] presented a model that learns to copy words from input to output, focusing on the small difference between input and output in the sentence proofreading task. In this study, we adopt Gu et al.'s model as the back translation model. We use difficult Japanese sentences in "SNOW T15: Japanese Simplified Corpus with Core Vocabulary" as input and learns simplified Japanese as output. We input pseudo-post-proofread sentences into the learned model and generate pseudo-pre-proofread sentences.

H. Generating Pseudo-pre-proofread Sentences

The goal of this study is to improve the proofreading accuracy from the baseline, seq2seq, which is constructed from GRU. Since the extended dataset is generated from the input corpus, the model is expected to get overfitting. In our experiments, we apply word-dropout to seq2seq, which sets the vector of input words to zero with a certain probability.

TABLE III
EXPERIMENTAL RESULTS

No Expansion	Back Translation	Rule Based	Word2vec
11.92	12.42	12.97	13.56

IV. EXPERIMENTS

A. The Dataset in This Study

Throughout this study, we use "SNOW T15: Japanese Simplified Corpus with Core Vocabulary". This corpus contains 50,000 of three pairs: simplified Japanese, which is easy for foreigners to understand; difficult Japanese, which is difficult to understand; and their English translations. In this study, only the simplified Japanese and the difficult Japanese are used. We treat simplified Japanese as correct sentences and difficult Japanese as incorrect sentences.

B. Experimental setting

We train seq2seq by adding sentences expanded by Markov chains and three different methods to the 40,000 sentences corpus that excludes the 10,000 sentences used to train the classifier. The seq2seq we want to improve consists of one layer of bi-directional GRUs. Dropout is applied to each layer, and their probabilities are set to 0.2. In the output layer, Softmax is used to calculate the probability of word occurrence. We use Sentencepiece [12] to perform morphological analysis on the input text data. Sentencepiece is capable of performing morphological analysis by building a dictionary to keep the corpus to a specified number of words by unsupervised learning. It eliminates the need to process low-frequency words in the corpus treated as out-of-vocabulary.

The classifier for sorting pseudo-post-proofread sentences is bidirectional, the size of the hidden layer is 128, dropout is applied to each layer, and its probability is 0.2. The activation function is set to sigmoid. The training of seq2seq is 50 epochs, the batch size is 128, the optimization function is Adam, and the loss function is Pytorch's NLLoss.

C. Experimental Results

In this study, 40,000 sentences were generated by Markov chains. Of these, there are no sentences in which all words overlap with those in the original corpus. Of the 40,000 sentences, 9,039 were deleted by the classifier because they were judged to be sentences that appeared to have been generated by the Markov chain. For the remaining pseudo-post-proofread sentences, we generate three types of pseudo-pre-proofread sentences using each of the three methods described in Section III.D.

Of the 40,000 entries in "SNOW T15: Japanese Simplified Corpus with Core Vocabulary" that were not used for training the classifier, 10,000 were used as the test data for seq2seq, and the remaining data were divided into training data and test data at a ratio of 8:2. To the training data, we added the pseudo-post-proofread sentences and pseudo-pre-proofread

ISBN: 978-988-14049-2-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

sentences generated by the Markov chain. We measure the accuracy of the trained seq2seq using the pre-proofread sentences of the test data as input. The results are shown in Table 3. The accuracy is measured by BLEU, which measures the accuracy by the degree of agreement of n-grams. The results are displayed as 0 to 1 and can be expressed as a percentage by multiplying the number by 100. In this study, the BLEU results are multiplied by 100 and rounded to the nearest hundredth.

V. Conclusion

This study showed that the Markov chain model and the three pseudo-pre-proofread sentence generation methods slightly improved the BLEU score of the seq2seq model in the proofreading task. This method is useful when the number of sentences in corpora is extremely small because it can increase the number of sentences in corpora both before and after proofreading. This is especially useful for proofreading tasks, where it is difficult to obtain the corpus before and after proofreading.

In the experimental results, even the most accurate one is only 1.64 percent improvement compared to the unextended one. There are two possible reasons for this. The first is overfitting. Since Markov chain models generate sentences by learning the word order probabilities of the sentences, they cannot output anything other than the words and word order that appear in the original corpus. As a result, a large number of sentences that are not full matches but partial matches will be generated. As a result, the training of the model may be hindered. In future research, it is necessary to devise a method for generating completely new pseudo-post-proofread sentences that do not violate the grammatical correctness of the words and word order of the post-proofread sentences. The second issue is the semantic correctness of the sentences generated by the Markov chain model. The accuracy of the GRU itself is not as high as expected, and the semantic correctness of the selected sentences cannot be expected. In the future, we need to change the classifier to a more accurate model.

References

[1] Japan National Tourism Organization（2020）: Tukibetsu・Nenbetsu Toukei-data(Hounichi Gaikokuzin・Syukkoku Nihonzin）(in Japanese)
https://www.jnto.go.jp/jpn/statistics/visitor_trends/

[2] Japan Tourism Agency（2020）: Hounichigaikokuzin ga Ryokouchu ni Komattakoto, Ukeirekankyouseibi no Kadai ga Akiraka ni narimashita ～ Ukeirekankyou ni tsuite Hounichigaikokuzinryokousya ni Anke-to tyousa wo zisshi~(in Japanese)
https://www.mlit.go.jp/kankocho/news08_000267.html

[3] Akihiko Sugimoto（2018）: RECRUIT no Kouetsu-AI ga Kyouiteki na Kouka Kenshuturitsu ha Hito wo koe Suubyou de Kanryou(in Japanese),NIKKEI XTREND.

[4] Hiromitsu Ota（2017）: A Study on Main Methods and Uniqueness in Automatic Text Generation, IPSJ SIG Technical Reports, Vol.2017-IFAT-127 No.3.

[5] Yuta Hitomi, Hideaki Tamori, Naoaki Okazaki, Kentaro Inui（2017）: Proofread Sentence Generation as Multi-Task Learning with Editing Operation Prediction, Proceedings of the Eighth International Joint Conference on Natural Language Processing, 436–441.

[6] Hiroto Nakazima, Tuyoshi Nakazima（2018）: Ayamaribun no Zidouseisei niyoru Kouseienzin no Gakusyuu, The Association for Natural Language Processing, NLP2018.

[7] Kazuhide Yamamoto and Takumi Maruyama（2018）: Simplified Corpus with Core Vocabulary. The 11th International Conference on Language Resources and Evaluation (LREC 2018).

[8] Pre-trained word vectors of 30+ languages（2020）– GitHub, https://github.com/Kyubyong/wordvectors

[9] Toshinori Sato：mecab-ipadic-NEologd.
https://github.com/neologd/mecab-ipadic-neologd

[10] Youichiro Ogawa, Kazunori Yamamoto（2020）: Nihongo Ayamari Teisei ni okeru Gizi Ayamari seisei ni yoru Kunren De-ta Kakutyou, Proceedings of the Twenty-sixth Annual Meeting of the Association for Natural Language Processing, pp.505-508.

[11] Jiatao Gu, Zhengdong Lu, Hang Li, Victor O.K. Li, “Incorporating copying mechanism in sequence-to-sequence learning”, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp.1631‐1640, 2016.

[12] SentencePiece‐GitHub, https://github.com/google/sentencepiece

ISBN: 978-988-14049-2-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)