Grammatical Error Correction in Japanese with Consideration of Error Types

2310067 SHI Haoda

Grammatical Error Correction (GEC) is a method that automatically corrects grammatical errors in texts written by language learners, which is widely applied for various fields such as language education. While research on GEC has primarily focused on English, studies on Japanese GEC remain limited. In addition, many previous studies have trained a single GEC model without considering error types, thus failing to fully account for the distinct characteristics of each type of error. Furthermore, it is crucial not only to correct grammatical errors but also to provide feedback on the nature of the errors in order to support language learners in understanding their own errors and acquiring correct grammar. However, such a research direction is not well studied in the field of Japanese GEC.

This study proposes a method for Japanese GEC that integrates an error type classification model and individual GEC models that are specialized for each of the error types. First, error types are predefined, and specialized GEC models are trained for each of the error types. For a given erroneous sentence, its type of error is identified. Then, the GEC model corresponding to the identified error type is applied to generate a corrected sentence. It is expected to improve the performance of GEC by utilizing specialized GEC models tailored to each error type. Finally, both the corrected sentence and the identified error type are presented to a learner. Showing the error type helps learners understand the cause of their error, facilitating more effective language learning.

The details of the proposed method are explained as follows. First, seven error types are defined: "postposition," "auxiliary verb," "typographical error," "verb conjugation," "incorrect choice of verb," "incorrect choice of noun," and "other." Pairs of erroneous sentences written by Japanese learners and sentences corrected by native Japanese speakers are extracted from the Lang-8 corpus, an existing dataset of GEC, with preprocessing of removal of edit markers and so on. Next, error types are assigned to each pair of erroneous and corrected sentences by the following procedures: detecting differences between an erroneous sentence and a corrected sentence, performing morphological analysis for word segmentation and part-of-speech tagging, and identifying an error type by manually designed rules. A pair of sentences is removed if an erroneous sentence contains multiple errors. Through these procedures, a dataset consisting of triplets of an erroneous sentence, a corrected sentence, and a label of the error type is constructed, which is then referred to as the "Type-labeled GEC dataset."

Next, the error type of an erroneous sentence is classified. Pre-trained learning models, specifically BERT and RoBERTa, are used as the base model. They are fine-tuned using the Type-labeled GEC dataset to train error type classification models. Furthermore, data augmentation is applied to increase the number of samples for "typographical error" and "incorrect choice of noun," as there are a few samples of these error types in the dataset. Grammatically correct Japanese sentences are converted to erroneous sentences by artificially introducing errors of "typographical error" or "incorrect choice of noun", where the KeiCo corpus serves as the source for grammatically correct sentences. These synthesized erroneous sentences are then coupled with their original sentences to create additional error correction samples. The generated pseudo-samples are combined with the original dataset, and the error type classification model is fine-tuned using this expanded training data.

Next, the method to revise an erroneous sentence into a grammatically correct sentence is implemented as follows. A sequence-to-sequence model is trained as a GEC model where an erroneous sentence is an input and a corrected sentence is an output. The pre-trained language model T5, which is applicable for sequence-to-sequence tasks, is fine-tuned using the Type-labeled GEC dataset. The dataset is divided into several portions by the error type, and a separate GEC model that focuses on one error type is trained using each portion of the dataset. As a result, seven different GEC models that are particularly tuned on the correction of the specific error type are obtained.

We report the experiments to evaluate our proposed method as follows. The error type classification models are evaluated on two tasks: the error classification task as well as the error detection and classification task. The former is a task to classify an error type of a given erroneous sentence into seven classes (error types). The latter is a task to identify whether a given sentence contains a grammatical error and classify its error type if there is an error, implemented by adding "no error" as a classification class. In the second task, grammatically correct sentences extracted from the Lang-8 corpus are labeled with the "no error" tag and added to the dataset. A dataset consisting of 150 erroneous sentences manually labeled with their error types is prepared and used as test data to evaluate the performance of the error type classification models.

In the error classification task, a comparison between the two language models used as the base model for error type classification showed that RoBERTa achieved a higher F1 score than BERT. The data augmentation did not improve the F1 score for BERT, but it led to an improvement for RoBERTa. The highest F1 score was 0.60, which was obtained by the

RoBERTa model with data augmentation. On the other hand, in the error detection and classification task, BERT outperformed RoBERTa, achieving an F1 score of 0.55.

Next, the performance of the GEC models is evaluated. A single GEC model that corrects errors without taking error types into account is used as a baseline, and its performance is compared with the proposed method where individual GEC models for each of the error types are employed. Two variations of the proposed method are considered: $\mathbf{PRO_{gold}}$, which switches GEC models based on the ground-truth error type in the dataset, and $\mathbf{PRO_{auto}}$, which switches the models based on the automatically classified error type. In the $\mathbf{PRO_{auto}}$, the error type classification model is the RoBERTa model that has been fine-tuned with the augmented dataset. The GLEU score is used as the evaluation criterion for GEC. $\mathbf{PRO_{gold}}$ outperformed the baseline in all error types except for "postposition." The overall GLEU score of $\mathbf{PRO_{gold}}$ (0.7739) was higher than that of the baseline (0.7618). However, the GLEU score of $\mathbf{PRO_{auto}}$ was 0.7390, which was lower than the baseline. Especially, a significant decline in performance was observed for the "auxiliary verb" error type. This may be caused by the low performance of error type classification. The improvement of the performance of the error type classification model remains important future work.