Semi-supervised learning and bidirectional decoding for effective grammar correction in low-resource scenarios

ABSTRACT

The correction of grammatical errors in natural language processing is a crucial task as it aims to enhance the accuracy and intelligibility of written language. However, developing a grammatical error correction (GEC) framework for low-resource languages presents significant challenges due to the lack of available training data. This article proposes a novel GEC framework for low-resource languages, using Arabic as a case study. To generate more training data, we propose a semi-supervised confusion method called the equal distribution of synthetic errors (EDSE), which generates a wide range of parallel training data. Additionally, this article addresses two limitations of the classical seq2seq GEC model, which are unbalanced outputs due to the unidirectional decoder and exposure bias during inference. To overcome these limitations, we apply a knowledge distillation technique from neural machine translation. This method utilizes two decoders, a forward decoder right-to-left and a backward decoder left-to-right, and measures their agreement using Kullback-Leibler divergence as a regularization term. The experimental results on two benchmarks demonstrate that our proposed framework outperforms the Transformer baseline and two widely used bidirectional decoding techniques, namely asynchronous and synchronous bidirectional decoding. Furthermore, the proposed framework reported the highest F1 score, and generating synthetic data using the equal distribution technique for syntactic errors resulted in a significant improvement in performance. These findings demonstrate the effectiveness of the proposed framework for improving grammatical error correction for low-resource languages, particularly for the Arabic language.