Abstract

Recently, the Inverse Text Normalization (ITN) problem has received significant attention when the Automatic Speech Recognition (ASR) system has applicability in the production environment. The task aims to convert spoken form text into the corresponding written form format. ITN plays an important role when it helps to improve the readability of the ASR's output and improve the efficiency of natural language processing tasks behind the ASR system.

This problem is challenging when it requires understanding the context to normalize the text correctly. Traditional methods based on grammar rules do not require training data but are less effective for cases with complex contexts. Developing a system based on grammar rules is often complicated due to language dependence and the need for language experts. This method is also difficult to apply to the two subproblems of ITN: Restoring punctuation and Capitalizing proper nouns, because they require deep context understanding. Recent studies using deep learning proved to be quite effective, and system development also becomes more accessible when it does not depend on language or grammar rules. This method also overcomes the limitation of the traditional approach when the input has a complex context. But deep learning requires many data to label the training process, which is expensive. In addition, studies often focus on the problem of converting text from spoken form to written form, ignoring two sub-problems: restoring punctuation marks and capitalizing proper nouns. The deep learning method can generate unrecoverable errors if the input sentence contains rare words. However, research based on deep learning gives results superior to traditional methods. However, it still has limitations: (1) Using a lot of labeled data, ignoring punctuation and capitalization restoration tasks, and (3) the model often generates unrecoverable errors.

To overcome the above limitations of the deep learning method, we propose to use a deep language model that is trained first and then refined for the ITN problem. Using a pre-trained language model makes it possible for the model to use available contextual information instead of learning from scratch. Pre-trained helps the ITN need very little labeled data compared to training from scratch. However, using less training data means the unrecoverable errors appear more. For this problem, we propose to apply the subword technique to separate rate words into smaller units before training the model. The subword method can handle the rare word problem in ITN thoroughly. However, using fewer data makes more unrecoverable errors. For

this problem, we propose to apply the subword technique to separate rate words into smaller units before training the model. Besides, we also train our model simultaneously to process two additional problems, including restoring punctuation marks and capitalizing proper nouns.

The experimental results of the proposed method are much higher than the baseline method. Above all, the experimental results show that the model achieves good results for two subproblems: Restoring punctuation marks and Capitalizing proper nouns. For English data, our proposed method gives competitive results with the state-of-the-art. Additionally, our solution using subwords shows significantly improved results than word units on both English and Vietnamese datasets.