A Pre-trained Language Model for Chinese Pinyin-to-Character Task Based on BERT

Zhang Yu, Inwhee Joe*
Hanyang University
zhangyu140402@gmail.com, *iwjoe@hanyang.ac.kr

Abstract

The pre-trained models such as BERT have been reached state-of-the-art results in many downstream tasks of natural language understanding. It is obvious that the pre-trained model can make better use of the feature representation of the context to build a robust language model. Hence, in this paper, we propose a novel BERT-P2C model which embedding the Chinese word vectors and character vectors as joint input to solve the downstream task of Chinese Pinyin-to-Character. The results show that BERT-P2C model has a significant improvement over other non-pretrained models and the original BERT.

I. Introduction

Pinyin is the official romanization representation of Chinese characters and a tool to assist the pronunciation of Chinese. Pinyin provides a solution that maps Chinese character to a string of Latin alphabets so that each character has a letter writing form of its own. Over 20000 different Chinese characters can be mapped into 500 Pinyin syllables. Although the commonest Chinese characters are only over 6000. This also means that one syllable may correspond to multiple Chinese characters and a character may also correspond multiple pinyin syllables. These quite common homophones and polyphones present complexity and ambiguity of Pinyin-to-Character conversion task. Conventionally, the solution of P2C task rely on a predefined dictionary and the results depend on the quality of the dictionary and lack of flexibility. Hence, it is natural to regard the P2C conversion as a kind of machine translation subtask and can be solved with a pre-training language model.

II. Related Work

As the publication of the attentional mechanism [1], pre-trained language models (PTMs) greatly improves the state of art on various natural language processing tasks with large text corpus. Generally, PTMs can be divided into three categories, autoregression model (AR), Denoising autoencoder model (DAE) and Permutated Language Model (PLM). The essence of AR model is the unbiased estimation of joint probability. It considers the correlation between the predicted words that is naturally suitable to handle the naturally generated task. However, the joint probability cannot obtain the representation of bidirectional context information by disassembling the text sequence from left to right. Representative models are ELMo[2], GPT-2, DAE models such as BERT[3], RoBERTa[4] construct masked language model (MLM) by introducing noise MASK to obtain contextual bidirectional feature representation. PML promotes the conventional autoregressive language model, transforms sequential disassembly and generates context-dependent bidirectional feature representations. According to the characteristics of P2C tasks, the bidirectional representation capability of the MLM model is more effective than the left-to-right approach of the AR model. Moreover, due to the open source data we can obtain and the limitations of the experimental conditions, we are temporarily unable to handle text-level tasks, so it is difficult to use XLNet[5] with text-level long-distance dependence. By comparing the characteristics of these models, we decided to use BERT as our main optimization target model and proposed a BERT-P2C model.

III. Methodology

As shown in Figure 1, our model is mainly composed of two parts, one is based on the optimize BERT pre-training model, and the other is the conventional transformer encoder-decoder model. The specific structure and function of the transformer model will not be explained in detail in this article. BERT pre-training model, and the other is the conventional transformer encoder-decoder model. The specific structure and function of the transformer model will not be explained in detail in this article.

1. Pre-training

In the pre-training process, we hope to take advantage of the efficient bidirectional representation and semantic

Fig 1. The main structure of the BERT-P2C model
reasoning capabilities of the BERT model, so we chose the BERT model to handle MLM tasks and trained with a large Chinese corpus. At the same time, we have also made some improvements to some shortcomings in the BERT model. Because the original text of BERT’s processing of Chinese is to directly treat each word as a token, this will cause some words with a fixed collocation relationship without contextual semantic inference is easily predicted, which reduces the effect of pre-training to a certain extent. To tackle this issue, we used the word segmentation toolkit THULAC [6] for the Chinese input sentences and then we input the word segmentation results into the segment embedding layer. Besides, in order to solve the out-of-vocabulary problem, the embedding of Chinese words is also necessary, so we integrated BERT’s original token embedding with segment embedding and input it into the self-attention layer.

For the masking strategy, we used a dynamic mask strategy that duplicated the training data ten times, each time randomly masking 15% of the token or segment.

2. Fine-tuning
To utilize the pre-trained knowledge, we froze the parameters in pre-trained BERT and integrated the BERT output with encoder and decoder token embedding. In fine-tuning stage, we used the down-stream data set for P2C tasks to train the whole BERT-P2C model, the introduction of the dataset will be in section 3.1. Since we have not found a tool that can segment Pinyin, the input of the BERT part during fine-tuning is the token and segment of Chinese text instead of pinyin. In addition, we only used a static masking strategy in the encoder-decoder part of the transformer model.

IV. Experiment
1. Dataset
We used the corpus of CLUECorpus2020 to pre-train the BERT model. CLUECorpus2020 has 100G raw corpus with 35 billion Chinese characters, which is retrieved from Common Crawl and by using this corpus for pre-training, SOTA results have been achieved in many downstream tasks.

For fine-tuning stage, we used AISHELL-1[7] and THCHS-30[8] two open-source Mandarin speech corpus which include more than 200 hours Mandarin speech recording text with clearly marked pinyin label as down-stream data. We have integrated the training set and test set of the two databases respectively. The training set approximately equal to 180h and 15h data as development set over 6h text data as test set.

2. Experiment settings
In the pre-training part, due to the limitation of our hardware, we followed the hyperparameter settings of the BERT base model which include 12 transformer blocks and 110m parameters. And the number of Multi-Head Attentions is 16, the hidden size is 768.

For fine-tuning part, in order to reduce part of the computation pressure, we set up three layers of transformer blocks in the encoder and decoder parts and set up batch size to 64. Optimized the model with Adam and the learning rate is 4e-4.

3. Experimental results
In order to verify the effectiveness of the model, we made comparisons with ELMo model and Google BERT base model. According to the results in Table 1, compared with the ELMo model, the BERT-based pre-trained model has a significant improvement in results. At the same time, the model BERT-P2C based on the BERT model for the Pinyin-to-character task improved the results by 3.1% of accuracy compared with the original model in both of development set and test set.

<table>
<thead>
<tr>
<th>Dev. (acc%)</th>
<th>Test (acc%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELMo</td>
<td>89.3</td>
</tr>
<tr>
<td>BERT_base</td>
<td>92.7</td>
</tr>
<tr>
<td>BERT-P2C</td>
<td>95.8</td>
</tr>
</tbody>
</table>

Table 1. Accuracy in development and test set comparing with different models.

This shows that our improvements to the BERT model, the design and training strategy of the entire model, and the hyperparameter selection are very effective.

V. Conclusion
In this paper, we proposed a new pre-trained language model BERT-P2C to deal with Chinese Pinyin-to-character down-stream task. Without further expanding the parameter scale, compared with the previous non-pre-training model and the traditional BERT model, the model performance has been greatly improved.

References: