ICANN19
28th International Conference on Artificial Neural Networks

An Improved Method of Applying a Machine Translation Model to a Chinese Word Segmentation Task

Yuekun Wei, Binbin Qu*, Nan Hu, and Liu Han

Huazhong University of Science and Technology, Wuhan, China
Contact Author：Binbin Qu （binbinqu_hust@163.com）

ABSTRACT

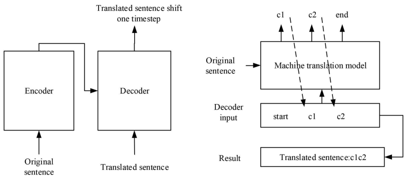
In this paper, we propose a novel method named Translation Correcting to solve the problem that applying the Machine Translation (MT) model to Chinese Word Segmentation (CWS) task would introduce translation errors and get a new model named CWSTransformer. Translation Correcting eliminates translation errors by utilizing the information of a sentence that needs to be segmented during the translation process. Consequently, the performance of word segmentation is considerably improved.

.

INTRODUCTION

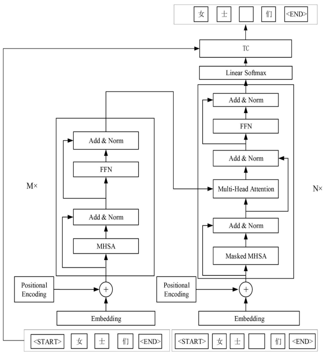
Chinese Word Segmentation (CWS) has been widely researched recently, and most commonly used method is to treat CWS as a sequence labeling problem, or consider the context within a fixed window using deep neural networks.

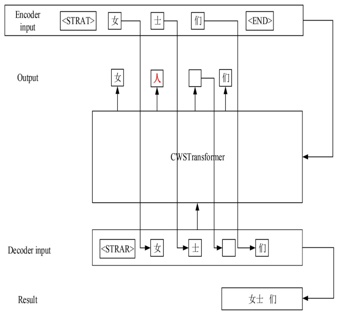
Recently, A new idea to treat CWS as a Machine Translation (MT) task has been proposed with structure in the figure. Because of the characteristic of MT, during the process of translation, it will make the determination of the next predicted character unclear and result in an incorrect character. To solve this problem, some researchers utilize a MT model with post-editing method to solve CWS. This method can only correct some incorrect translated characters but not incorrect segmentation results. we propose a novel method named Translation Correcting to correct the translation results at each step and obtain a new model CWSTransformer by improving the MT model with the new method.



METHODOLOGY

Using Translation Correcting to improve the MT model Transformer, we obtain a new CWS model, CWSTransformer. The architecture of the CWSTransformer is shown in the left Figure. This model is based on an encoder-decoder framework, where both the encoder and decoder are composed of stacked blocks.





The translation process of CWSTransformer is shown in the right Figure. Translated sentence is generated step by step. The model uses the generated sentence to predict the next translated character at each step and the translation correcting method to correct the error.

The details of the translation process can be formalized into algorithm.

In the algorithm, steps 4 to 10 show the translation process. After we select the character with the highest probability as the result of translation $last\_token$ at step 6. To correct the result, if the $last\_token$ is not segmentation character $index_{seg}$, we set the $last\_token$ to the character at pos in S_{orig} and move pos one step forward. After that, we append $last\_token$ to S_{seg}. We repeat the translation process until we encounter the end symbol and get the result of CWS.

.

```
Algorithm 1: Translation Correcting
Input  : $model$: Machine translation model;
         $S_{orig}$: An original character sequence represented in index with a
         start and end symbol index;
         $index_{seg}$: Segmentation symbol index in dictionary;
         $index_s$: Start symbol index in dictionary ;
         $index_e$: End symbol index in dictionary ;
Output : $S_{seg}$: Corresponding segmented character index sequence;
1  Initialize empty set $S_{seg}$
2  Append $index_s$ to $S_{seg}$
3  Initialize position $pos$ : $pos \leftarrow 1$
4  while $S_{orig}^{pos} \neq index_e$ do
5      Predicting the probability distribution of the output character $predict \leftarrow$
         $model(S_{orig}, S_{seg})$
6      $last\_token \leftarrow argmax(predict)$
7      if $last\_token \neq index_{seg}$ then
8          $last\_token \leftarrow S_{orig}^{pos}$
9          $pos \leftarrow pos + 1$
10     Append $last\_token$ to $S_{seg}$
11 return $S_{seg}$
```

RESULTS && DISCUSSION

To evaluate the performance of CWSTransformer, we use the performance of the CWS tool Jieba on the benchmark datasets as the baseline. Then we compare the performance of previous translation-based CWS model, Transformer-1 (1 encoder block, 1 decoder block), CWSTransformer-1 (1 encoder block, 1 decoder block), and CWSTransformer-3 (3 encoder blocks, 3 decoder blocks) on the PKU and MSR datasets. We show the datasets and the result of scoring below.

Dataset	**Train**	**Development**
PKU	132368	14708
MSR	282700	31412

Models	PKU			MSR		
	Recall	Precision	F-Score	Recall	Precision	F-Score
Baseline	78.7	85.3	81.8	81.2	81.7	81.5
Cheng et al. (2017)	88.6	87.0	87.8	93.2	95.1	94.1
Transformer-1	88.8	87.8	88.3	88.9	85.9	87.4
CWSTransformer-1	89.2	90.0	89.6	88.9	89.9	89.5
CWSTransformer-3	91.7	92.3	**92.0**	94.5	94.2	**94.4**

The experimental results show that the performance of CWSTransformer is superior to Transformer and previous translation-based CWS model. By observing the correction results of post-editing in the referenced model and Translation Correcting on PKU dataset, we find that Translation Correcting can not only correct the translation errors of characters, but also correct the segmentation errors. An example is shown in the Table.

Origin sentence	... 向多个国家领导人打电话求援, ...
Correct result	... 向 多 个 国家 领导人 打电话 求援 , ...
Transformer	... 向 多 个 国家 领导人 打电话 话 援援 , ...
Post-editing	... 向 多 个 国家 领导人 打电求援 , ...
Translation Correcting	... 向 多 个 国家 领导人 打电话 求援 , ...

This verifies the validity of Translation Correcting.

Experiments also show that the increase in the number of encoder and decoder blocks improves the segmentation performance of CWSTransformer. Limited by resources, we could only evaluate a version of CWSTransformer containing 3 encoder and 3 decoder blocks.

CONCLUSION

This paper proposes an effective method for applying the MT model to a CWS task. This method can be applied to any MT model. Using this method, we improve the MT model Transformer and obtain a new CWS model, CWSTransformer. The experimental results show that CWSTransformer obtained by Translation Correcting corrects some translation errors in Transformer, thereby improving the word segmentation. CWSTransformer outperforms the CWS models proposed in previous studies. However, because of limited resources, we did not use a version of CWSTransformer containing more encoder and decoder blocks to conduct the experiments; thus, the performance of the model is limited in the experiment. In future works, we will try to improve CWSTransformer, which can use fewer resources to achieve better results.

Other contact Information:

Name: Yuekun Wei
Email: M201773002@hust.edu.cn
WeChat：WHK_14

ManuscriptID

295