Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 159–163
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022

Simple Tagging System with RoBERTa for Ancient Chinese

Binghao Tang, Boda Lin, Si Li*
School of Artificial Intelligence
Beijing University of Post and Telecommunication, China
{tangbinghao, linboda, lisi}@bupt.edu.cn

Abstract

This paper describes the system submitted for the EvaHan 2022 Shared Task on word segmentation and part-of-speech tagging for Ancient Chinese. Our system is based on the pre-trained language model SIKU-RoBERTa and the simple tagging layers. Our system significantly outperforms the official baselines in the released test sets and shows the effectiveness.

Keywords: EvaHan 2022, Word Segmentation, POS tagging

1. Introduction

Chinese Word Segmentation (CWS) is a fundamental task in Natural Language Processing (NLP). Generally speaking, word is the basic unit containing complete semantic information. Thus CWS is widely used for difference NLP tasks, such as machine translation (Yang et al., 2018), text classification (Zeng et al., 2018), and question answering (Liu et al., 2018). Comparing with CWS, Part-of-speech (POS) tagging is a more general task for many languages, which aims to assign pre-defined syntactical property for each token in the sentence. Some research (Ng and Low, 2004) validates combining them into a joint task can provide better performance than separately conducting these two tasks in a sequence. Thus the CWS is usually implemented with the prediction of POS tagging jointly in the recent years (Tian et al., 2020a).

Previous studies about this joint task are usually deemed as sequence labeling task (Zhang et al., 2016; Higashiyama et al., 2019; Qiu et al., 2020). These models achieve excellent performance in this task, especially with the wide usage of pre-trained language model (PLM) (Tian et al., 2020b). However, most Chinese versions of PLMs are pre-trained on the multilingual corpus (Devlin et al., 2019; Liu et al., 2020; Xue et al., 2021). Even if there are few PLMs pre-trained on pure Chinese corpus, most of them use the Morden Chinese (Sun et al., 2019). Recently there are some work release PLMs pre-trained on ancient Chinese corpus to withdraw this lacking in CWS, e.g., *SIKU-BERT* (Wang et al., 2021) and *SIKU-RoBERTa* (Wang et al., 2021). Based on these two models, the first NLP tool evaluation competition in the field of ancient Chinese, i.e, EvaHan 2022 is released. EvaHan 2022 aims to exploit an efficient way to handle the joint task of CWS and POS tagging on ancient Chinese language.

In this paper, we describe our submitted system for the EvaHan 2022. Our system is based on the released ancient Chinese version of RoBERTa (Wang et al., 2021). We utilize extra knowledge from ancient Chinese via the pre-trained RoBERTa, and further encode features by concrete context information with Bi-LSTMs.

The experimental results on the two test sets demonstrate the effectiveness of our method. Our method significantly outperforms the official baselines to a large margin in the in-domain test set.

2. Related Work

2.1. Chinese Word Segmentation & POS tagging

Chinese Word Segmentation (CWS) has been studied for a long time, as one of the most fundamental NLP tasks for Chinese language processing (Higashiyama et al., 2019; Qiu et al., 2020). And part-of-speech (POS) tagging is also a basic task for natural language processing. Some research (Ng and Low, 2004) demonstrates that combining CWS and POS tagging tasks together as a joint task can improve both of them. So many researchers dedicate to CWS and POS tagging and obtain many amazing achievements (Tian et al., 2020a). However, most research is based on modern Chinese while few works pay attention to ancient Chinese. Considering this situation, EvaHan 2022 release a competition for the joint task on ancient Chinese.

2.2. Pre-trained Language Model

BERT (Devlin et al., 2019) is widely used PLM for CWS (Tian et al., 2020a). Besides, (Wang et al., 2021) apply RoBERTa to implement CWS task. while there are many differences between modern Chinese and ancient Chinese. So straightly using the PLMs in the area of ancient Chinese usually gets unsatisfactory performances. Thus *SIKU-RoBERTa* (Wang et al., 2021), which continues to train on ancient Chinese corpus based on vallina Chinese RoBERTa (Liu et al., 2019), seems to be a good choice.

3. Method

We introduce the overall procedure of our system for this evaluation task, which includes the pre-processing, model architecture and the solution for the long sentence.

*Corresponding author

https://circse.github.io/LT4HALA/2022/EvaHan

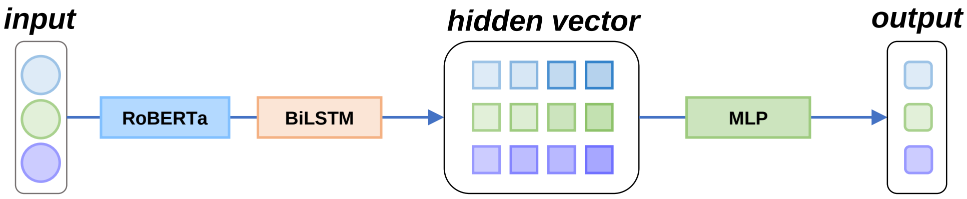

Figure 1: Overall Architecture

Set	Domain	Number of Word Tokens	Number of Character Tokens
Train	Zuozhuan	$166,142$	$194,955$
Test A	Zuozhuan	$28,131$	$33,298$
Blind Test B	Other ancient Chinese Book	Around $40,000$	Around $50,000$

Table 1: The statistics data of the datasets.

3.1. Pre-processing

We firstly pre-process raw data. For example, the input sentence is “春秋/n 左/n 定公/n”. To start with, we split sentence into single tokens and use notation to distinguish each token’s position in origin word, i.e., B short for Begin, M short for Mid, E short for End, and combine it with its POS label. So the processed sentence should be like: “春 b-n 秋 e-n 左 b-n e-n 定 b-nr 公 e-nr”.

3.2. Model

The architecture of our model is shown in Figure 1.
We define the input sequence is $S = \{c_1, c_2, ..., c_n\}$, where c_i is the i-th character of the input sentence. The input S is sent into the RoBERTa, a multi-layer Transformer (Vaswani et al., 2017) structure model. In the l-th layer of Transformer, the hidden representation H_l is calculated as following:

$$\hat{H}_l = LayerNorm(H_{l-1} + Attention(H_{l-1})) \quad (1)$$

$$H_l = LayerNorm(\hat{H}_l + FFN(\hat{H}_l)) \quad (2)$$

where the H_0 is S, $LayerNorm$ is the layer-wise normalization layer, and the $Attention$ is the multi-head attention layer. Please refer to the original paper (Devlin et al., 2018; Liu et al., 2019) for more details.
After obtaining the encoding representation H from RoBERTa, a bidirectional LSTM is applied to further encoding the context representation :

$$R = BiLSTM(H) \quad (3)$$

Finally, we use a Multi-layer Proc (MLP) to predict the labeling sequence :

$$Y = MLP(R) \quad (4)$$

3.3. Solution for Long Sentences

The dataset contains some long length sentences, which are beyond the maximum length processed by the proposed model. Considering this situation, we split these long sentences into some short sub-sentences. We try to keep all sub-sentences semantically complete thus we split the long sentence according to punctuation instead of the maximum length. Then we revert sentences from the output file of system and obtain our final submit file.

Hypermeter	Value
learning rate	$2 \times e^{-3}$
layer of BiLSTM	3
dimension of embedding	300
hidden dimension of BiLSTM	400
dimension of MLP	500
dropout ratio	0.33

Table 2: hypermeters

4. Experiments

4.1. dataset

We use the datasets released by the host of EvaHan 2022, which include one training set and two test sets. All the sentences are collected from the ancient Chinese texts like Zuozhuan (Li et al., 2012). The details about the statistics of the dataset are shown in Table 1. The training data contains punctuated, word-segmented and part-of-speech tagged text from zuozhuan, an ancient Chinese work. There are two test data sets. Test A contains different data from the same book of training data. And test A also have annotated version in the form of training data, so test A can be used as validation sets while training. Test B contains texts which have similar content from different books and only Chinese characters and punctuation. Thus test B is designed as out-of-domain sets to test the generalization of system.

4.2. Implementation Details

We use the RoBERTa as the backbone for all experiments. The PLM is implemented with Huggingface

Task (Test A)	CWS			POS		
	Precision	Recall	F1	Precision	Recall	F1
Baseline	90.64	92.08	91.35	89.06	89.54	89.30
*SIKU-BERT	–	–	88.84	–	–	90.10
*SIKU-RoBERTa	–	–	88.88	–	–	90.06
Our System	**95.81**	**96.52**	**96.16**	**90.90**	**91.57**	**91.24**

Task (Test B)	CWS			POS		
	Precision	Recall	F1	Precision	Recall	F1
Our System	94.04	90.59	92.28	86.86	83.67	85.24

Table 3: Experimental results on two tests in terms of F1 Score. The host of EvaHan 2022 do not report the results of official baseline and results of SIKU-BERT, and SIKU-RoBERTa on Test B. * means the results about POS tagging of these two models are different from joint task of CWS and POS tagging.

Transformers . We use different learning rates for PLM and non-PLM layers in the model. The learning rate for PLM is $5 \times 10e^{-5}$, and the learning rate for non-PLM layers is $2 \times 10e^{-3}$. The optimizer is Adam (Kingma and Ba, 2014). We implement all experiments on Nvidia GTX1080Ti. Our system consumes about 7GiB GPU memory and it takes about 4 hours to achieve the best performance.

The other important hyperparameters are listed in Table 2.

Instead of performing a hyperparameter search, we directly chose the values of the parameters empirically.

4.3. Metric

Following the convention of CWS and POS tagging, we use Precision (P), Recall (R), and F1 Score as the evaluation metrics for all experiments. All the results are presented in percentages (%).

4.4. Baselines

We compare our system with the official baselines, which obtains on *Zuozhuan_test* using Conditional Random Fields (CRF) training on *Zuozhuan_train* without additional resources (Xiao-he, 2010). Besides, we also choose the BERT and RoBERTa pre-trained on the SIKU, which are noted as *SIKU-BERT* and *SIKU-RoBERTa* in Table 3.

4.5. Results

The results are shown in Table 3. Our system outperforms all the baselines in all metrics on both the CWS task and the POS tagging task, which demonstrate the effectiveness of our system. Besides, our system also obtain better performance comparing with the vanilla SIKU-RoBERTa. This comparison also can be regard as a ablation study, which validate the effectiveness of the additional layers we designed.

Algorithm 1: post-process

```
input tokens: $w_1/y_1, w_2/y-2, w_3/y_3, w_4/y_4$
output       : $w_1w_2w_3 \quad w_4$
for $i \leftarrow 1$ to $N$ do
    if $y_i = b$ and $y_{i+1} = b$ then
        $w_i \quad w_{i+1}$
    if $y_i = b$ and $y_{i+1} = s$ then
        $w_i \quad w_{i+1}$
    if $y_i = m$ and $y_{i+1} = b$ then
        $w_{i-1}w_i \quad w_{i+1}$
    if $y_i = m$ and $y_{i+1} = s$ then
        $w_{i-1}w_i \quad w_{i+1}$
    if $y_i = e$ and $y_{i+1} = e$ then
        $w_{i-1}w_iw_{i+1}$
    if $y_i = e$ and $y_{i+1} = m$ then
        $w_{i-1}w_i \quad w_{i+1}$
```

4.6. The Legality

The legality is an important issue for CWS task. The neural network may predicts some illegal labeling tokens such as "$w_1/b \quad w_2/b$". A traditional approach to dealing with this problem is using CRF to constrain the output sequence (Xiao-he, 2010). We do not apply CRF in our system for brevity, and the statistics results show only the 0.6% tokens in the test set are illegal.

For those illegal tokens, we correct them by post-processing which is shown in Algorithm 1. This low illegal ratio demonstrates that the great learning ability of RoBERTa can enables the model to learn implicit constraints between output labels (Liu et al., 2019).

5. Conclusion

In this paper, we describe the simple tagging system submitted for the EvaHan2022. The proposed system apply a pre-trained RoBERTa and the BiLSTM layers to encoding context information. The experimental results on the official test sets demonstrate the effectiveness of our system, especially the comparison between our system and the original official RoBERTa validate

https://huggingface.co/SIKU-BERT/sikuroberta the effectiveness of the additional tagging layers. Besides, we also discuss the legality issue for CWS.

6. References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Higashiyama, S., Utiyama, M., Sumita, E., Ideuchi, M., Oida, Y., Sakamoto, Y., and Okada, I. (2019). Incorporating word attention into character-based word segmentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2699–2709, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, B., Xi, N., Feng, M., and Chen, X. (2012). Corpus-based statistics of pre-qin chinese. pages 145–153, 07.

Liu, Z., Peng, E., Yan, S., Li, G., and Hao, T. (2018). T-know: a knowledge graph-based question answering and infor-mation retrieval system for traditional Chinese medicine. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 15–19, Santa Fe, New Mexico, August. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ng, H. T. and Low, J. K. (2004). Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 277–284, Barcelona, Spain, July. Association for Computational Linguistics.

Qiu, X., Pei, H., Yan, H., and Huang, X. (2020). A concise model for multi-criteria Chinese word segmentation with transformer encoder. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2887–2897, Online, November. Association for Computational Linguistics.

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., and Wu, H. (2019). Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Tian, Y., Song, Y., Ao, X., Xia, F., Quan, X., Zhang, T., and Wang, Y. (2020a). Joint Chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8286–8296, Online, July. Association for Computational Linguistics.

Tian, Y., Song, Y., Xia, F., Zhang, T., and Wang, Y. (2020b). Improving Chinese word segmentation with wordhood memory networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285, Online, July. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, D., Liu, C., Zhu, Z., Jiang, Feng, Hu, H., Shen, S., and Li, B.-S. (2021). Construction and application of pre-training model of “siku quanshu” oriented to digital humanities.

Xiao-he, C. (2010). Crf based research on a unified approach to word segmentation and pos tagging for pre-qin chinese. *Journal of Chinese information processing*.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

Yang, Z., Chen, W., Wang, F., and Xu, B. (2018). Improving neural machine translation with conditional sequence generative adversarial nets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1346–1355, New Orleans, Louisiana, June. Association for Computational Linguistics.

Zeng, J., Li, J., Song, Y., Gao, C., Lyu, M. R., and King, I. (2018). Topic memory networks for short text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing, pages 3120–3131, Brussels, Belgium, October-November. Association for Computational Linguistics.

Zhang, M., Zhang, Y., and Fu, G. (2016). Transition-based neural word segmentation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 421–431, Berlin, Germany, August. Association for Computational Linguistics.