Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 146–149
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022

Ancient Chinese Word Segmentation and Part-of-Speech Tagging Using Data Augmentation

Yanzhi Tian, Yuhang Guo*
School of Computer Science, Beijing Institute of Technology, Beijing 100081, China
{tianyanzhi, guoyuhang}@bit.edu.cn

Abstract

We attended the EvaHan2022 ancient Chinese word segmentation and Part-of-Speech (POS) tagging evaluation. We regard the Chinese word segmentation and POS tagging as sequence tagging tasks. Our system is based on a BERT-BiLSTM-CRF model which is trained on the data provided by the EvaHan2022 evaluation. Besides, we also employ data augmentation techniques to enhance the performance of our model. On the Test A and Test B of the evaluation, the F_1 scores of our system achieve 94.73% and 90.93% for the word segmentation, 89.19% and 83.48% for the POS tagging.

Keywords: ancient Chinese, word segmentation, POS tagging, data augmentation

1. Introduction

Ancient Chinese (a.k.a. classical Chinese) is a written language of Chinese used widely around 1000 BC to 221 BC. Most of the ancient Chinese records are written in classical Chinese. The classical Chinese is different from modern Chinese in several aspects, including wording and syntax. In order to study ancient Chinese automatically, classical Chinese word segmentation and Part-of-Speech (POS) tagging are of high research values.

Compared with the research on word segmentation and POS tagging of modern Chinese, the corpus of ancient Chinese with label is insufficient. The evaluation of EvaHan2022 provides a set of labeled corpus selected from the Zuozhuan corpus (Chen, Xiaohe, et al., 2017) and a pre-trained model called SikuBERT (Wang et al., 2021) which is trained based on ancient Chinese corpus.

We build an end-to-end ancient Chinese word segmentation and POS tagging system based on SikuBERT and attended the EvaHan2022 evaluation. We train our model on the given corpus. To ease the shortage of the labeled corpus, we employ data augmentation techniques. On the Test A and Test B of the evaluation, the F_1 scores of our system achieves 94.73% and 90.93% on the word segmentation, 89.19% and 83.48% on the POS tagging.

Our codes and results are available at `https://github.com/YanzhiTian/EvaHan-2022`.

2. Method

2.1. Model

We regard the word segmentation and POS tagging as sequence tagging tasks. BiLSTM-CRF is a well known sequence tagging model, in which the BiLSTM layers utilize both past and future input features efficiently, and the CRF layer reduces the possibility of the appearance of the illogical output tagging sequence (Huang et al., 2015). BERT(Devlin et al., 2018) is a pre-trained model and it is proved that the fine-tuning BERT-CRF model performances well on NER which is also a sequence tagging task (Souza et al., 2019). Here we apply the BERT-BiLSTM-CRF model.

In our system, we use the final output of SikuBERT as the input of the BiLSTM layer. We use dropout (Srivastava et al., 2014) to avoid overfitting and a linear layer to project the BiLSTM features to a lower dimension which corresponds to the input of CRF layer. The architecture of our model is shown in Figure 1.

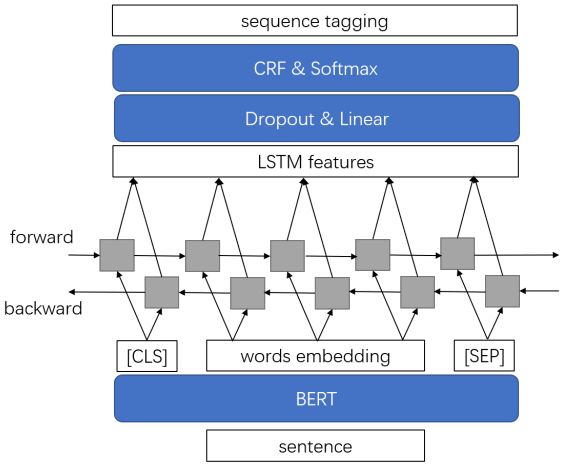

Figure 1: The architecture of our model.

2.2. Data

Our end-to-end model jointly handles the word segmentation and POS tagging which avoids the error propagation in the cascade model. We design a series of taggings including both word segmentation information and POS information. For example, the tagging

* Corresponding author.

"b-n" refers to the beginning of a word segment and noun POS, the tagging "i-n" refers to the middle or end of a word segment and noun POS. We have 47 kinds of tagging (including [PAD], [CLS] and [SEP]) in total.
An example of tagging is shown in Figure 2, the first row is the raw sentence, the second row is the tagging sequence in raw training set and the third row is the tagging sequence after processing.

宋	武	公	生	仲	子	。
/nr			/v	/nr		/w
/b-nr	/i-nr	/i-nr	/b-v	/b-nr	/i-nr	/b-w

Figure 2: An example of tagging.

2.3. Data Augmentation

We use data augmentation to ease the shortage of the labeled data and to enhance the performance of the model. Our strategy of data augmentation is to mask several words with a special token [MASK] dynamically. Before the sequence is input into the model, our system will generate a boolean array randomly to mask the words in the sequence.
Our motivation is that the model predicts tagging sequence harder compared with modern Chinese because a specific word in ancient Chinese are more variation of semantics corresponding to different kinds of POS tagging. Using our data augmentation method, the model can inference taggings from other taggings in the context instead of its word token which means the model can learn information from the sentence structure such as the sequence of POS tagging.
The mask rate should be chosen carefully. An appropriate mask rate will make the model has better performance. However a larger mask rate will reduce the performance of the model.

3. Experiments

We only use Zuozhuan_Train dataset which is provided by the EvaHan2022 evaluation to train our model. To evaluate the performance of our model, we shuffle the dataset and sample 900 sentences randomly to construct a validation set, the rest of the data to construct training set. The hyper-parameters of our model are shown in Table 1.

Hyper-parameter	Value
Learning rate	0.01
Batch size	64
Hidden dimension	2×512
LSTM layers	2
Dropout rate	0.5
Mask rate	0.2

Table 1: The hyper-parameters of our model.

The max sequence length of SikuBERT is 512(including [CLS] and [SEP]). We truncated the sentence by punctuation and kept the length of the sentence smaller than 512.

3.1. Training

In our system, the optimizer is Adam and the loss function is negative log likelihood calculated in the CRF layer. In the training step, we froze the parameters of BERT to make sure the error will not pass to the BERT layer in backpropagation because the size of our training set is much smaller compared with the size of the data used in the pre-training. This method can accelerate the convergence of model and make the training easier.

3.2. Ablation Study

We trained 4 models with different settings including BERT-Linear, BERT-CRF, BERT-BiLSTM-CRF and the Deeper Model. We also tested the mask rate of 0.2 and 0.3 on the BERT-BiLSTM-CRF model respectively. The results of these models on the validation set are shown in Table 2.
Compared with the BiLSTM-CRF model(Cheng et al., 2020), our BERT-BiLSTM-CRF model uses SikuBERT which is pre-trained on large scaled ancient Chinese corpus. We freeze the parameters of SikuBERT and use the final output as word embedding. The SikuBERT eases the shortage of the labeled corpus. Using data augmentation can introduce noises into data which is helpful to enhance the performance of the model and avoid overfitting.
The Deeper Model is a BERT-BiLSTM-Transformer Encoder-BiLSTM-CRF model. We evaluated the performance of the Deeper Model on the validation set in each epoch. The F_1 scores of the Deeper Model (solid lines) and BERT-BiLSTM-CRF model (dashed lines) of word segmentation and POS tagging in each training epoch are shown in Figure 3. The final F_1 scores of the Deeper Model are shown in Table 2.
It can be found that the F_1 scores of the Deeper Model get close to the final F_1 scores of BERT-BiLSTM-CRF model after about 50 epochs. However the BERT-BiLSTM-CRF model reaches the final F_1 score only after about 10 epochs which means the convergence of the Deeper Model is slower than BERT-BiLSTM-CRF model.
We evaluated the mask rate parameter with 0.2 and 0.3 on the validation set. As illustrated in Table 2, the evaluation results show that the mask rate with 0.2 performs better than 0.3. We use 0.2 as the mask rate parameter in our system.

4. Results

We evaluated our system on Test A and Test B closed modality tests of EvaHan2022 using BERT-BiLSTM-CRF model with data augmentation. The size and source of testing sets are shown in Table 3.

Model	WS			POS Tagging		
	P	R	F_1	P	R	F_1
BERT-Linear	93.08	92.97	93.04	84.97	84.87	84.92
BERT-CRF	93.13	93.02	93.08(+0.04)	85.27	85.17	85.22(+0.3)
BERT-BiLSTM-CRF (w/o DA)	94.24	94.49	94.37(+1.33)	88.42	88.65	**88.53(+3.61)**
BERT-BiLSTM-CRF (MR=0.2)	94.43	94.35	**94.39(+1.35)**	88.28	88.20	88.24(+3.32)
BERT-BiLSTM-CRF (MR=0.3)	93.60	94.36	93.98(+0.94)	87.12	87.84	87.48(+2.56)
The Deeper Model (MR=0.2)	94.40	94.30	94.35(+1.31)	87.52	87.42	87.47(+2.55)

Table 2: The precision(P), recall(R) and F_1 scores (%) of different models with different settings (without Data Augmentation (DA) and with different Mask Rates(MR)) on our validation set.

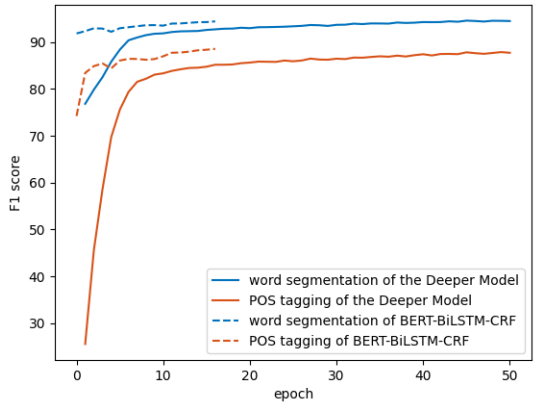

Figure 3: The F_1 scores (%) of BERT-BiLSTM-CRF model and the Deeper Model on validation set in each epochs.

Datasets	Sources	Word Tokens	Char Tokens
Test A	ZuoZhuan	28K	33K
Blind Test B	Other similar ancient Chinese Book	40K	50K

Table 3: The size and sources of test sets.

To verify the impact of data augmentation, we evaluated the performance of BERT-BiLSTM-CRF model without data augmentation. We also evaluated the performance of the Deeper Model to check the difference with other models. The results are shown in Table 4 and Table 5.

As shown in Table 4 and Table 5, the system with data augmentation achieves better performance on the POS tagging task: the F_1 scores are higher than the system without data augmentation by 1.79%, 2.41% on Test A and Test B respectively. However the effect of data augmentation for word segmentation is not significant. The system F_1 score is 1.05% higher than the system without data augmentation on Test A but is lower on Test B by 0.50%.

Compared with the widely used datasets on modern Chinese word segmentation and POS tagging, the size of the Zuozhuan(1.7M) dataset is similar to the size of PKU(1.1M) and MSRA(2.4M) dataset(Emerson, 2005) on word segmentation, however it is much smaller than the size of CTB5(4.9M) dataset(Xue et al., 2005) on POS tagging. So the improvement of data augmentation on POS tagging is more obviously than word segmentation.

Our detailed analysis shows that the most error of our system in the POS tagging comes from that our model can not distinguish the noun category including n, nr and ns representing common noun, person entity and location entity respectively.

The results also show that all the F_1 scores of the Deeper Model are lower than our system.

5. Conclusion and Future Work

In this paper, we implement an end-to-end ancient Chinese word segmentation and POS tagging system. We also propose a data augmentation method by masking words in the data using a special [MASK] token in this task. The results show that using data augmentation enhances the performance of BERT-BiLSTM-CRF model on ancient Chinese word segmentation and POS tagging. On Test A and Test B of testing data, our system achieves 94.73% and 90.93% F_1 scores on word segmentation, 89.19% and 83.48% F_1 scores on POS tagging.

In the future we plan to import an entity recognition module to improve hard POS taggings like n, nr and ns.

6. Bibliographical References

Cheng, N., Li, B., Xiao, L., Xu, C., Ge, S., Hao, X., and Feng, M. (2020). Integration of automatic sentence segmentation and lexical analysis of Ancient Chinese based on BiLSTM-CRF model. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 52–58, Marseille, France, May. European Language Resources Association (ELRA).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional

Model	Word Segmentation			POS Tagging		
	P	R	F_1	P	R	F_1
BERT-BiLSTM-CRF (w/o DA)	92.92	94.46	93.68	86.69	88.13	87.40
BERT-BiLSTM-CRF (**Our system**)	94.48	94.99	**94.73(+1.05)**	88.95	89.43	**89.19(+1.79)**
The Deeper Model	94.10	94.61	94.36(+0.68)	88.44	88.92	88.68(+1.28)

Table 4: The precision(P), recall(R) and F_1 scores (%) of our models on **Test A**.

Model	Word Segmentation			POS Tagging		
	P	R	F_1	P	R	F_1
BERT-BiLSTM-CRF (w/o DA)	92.79	90.13	**91.43**	82.27	79.91	81.07
BERT-BiLSTM-CRF (**Our system**)	93.07	88.89	90.93(-0.5)	85.45	81.61	**83.48(+2.41)**
The Deeper Model	88.49	89.60	89.03(-2.4)	80.36	81.38	80.86(-0.21)

Table 5: The precision(P), recall(R) and F_1 scores (%) of our models on **Test B**.

transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emerson, T. (2005). The second international Chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Souza, F., Nogueira, R., and Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

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.

Xue, N., Xia, F., Chiou, F.-D., and Palmer, M. (2005). The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

7. Language Resource References

Chen, Xiaohe, et al. (2017). *Ancient Chinese Corpus LDC2017T14*. Philadelphia: Linguistic Data Consortium, 1.0, ISLRN 924-985-704-453-5.