Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 150–154
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022

BERT 4EVER@EvaHan 2022: Ancient Chinese Word Segmentation and Part-of-Speech Tagging based on Adversarial Learning and Continual Pre-training

Hailin Zhang[1†], Ziyu Yang[1†], Yingwen Fu[1], Ruoyao Ding[2*]
Guangdong University of Foreign Studies, Guangzhou, China
[1]{20201010014, 20201002958, 20201010002}@gdufs.edu.cn,
[2]ruoyaoding@outlook.com

Abstract

With the development of artificial intelligence (AI) and digital humanities, ancient Chinese resources and language technology have also developed and grown, which have become an increasingly important part to the study of historiography and traditional Chinese culture. In order to promote the research on automatic analysis technology of ancient Chinese, we conduct various experiments on ancient Chinese word segmentation and part-of-speech (POS) tagging tasks for the EvaHan 2022 shared task. We model the word segmentation and POS tagging tasks jointly as a sequence tagging problem. In addition, we perform a series of training strategies based on the provided ancient Chinese pre-trained model to enhance the model performance. Concretely, we employ several augmentation strategies, including continual pre-training, adversarial training, and ensemble learning to alleviate the limited amount of training data and the imbalance between POS labels. Extensive experiments demonstrate that our proposed models achieve considerable performance on ancient Chinese word segmentation and POS tagging tasks.

Keywords: ancient Chinese, word segmentation, part-of-speech tagging, adversarial learning, continuing pre-training

1. Introduction

The Chinese nation has thousands of years of glorious history and culture with excellent cultural heritage mainly recorded through the ancient Chinese written language. It is a good choice to start with ancient classics if one wants to understand Chinese civilization and know about ancient Chinese literature, history, politics, economy, medicine, and other cultures. Classics can be said to be the inheritance carrier of Chinese civilization. Applying technologies such as big data and artificial intelligence (AI) to ancient books, digitizing them, and making them public can rejuvenate all dusty ancient books and make the words written in ancient books come alive, which can help more people know about the Chinese civilization. Research in the field of ancient Chinese is becoming more and more popular and important. For example, the national ancient book protection and digitization project has been listed as a key project in the inheritance and development of Chinese excellent traditional culture. This project not only speeds up the compilation and publication of ancient books but also facilitates knowledge extraction and information integration of ancient books and documents, providing a new method for the inheritance and protection of ancient books and injecting new vitality.

Although the automatic analysis of modern Chinese has achieved promising results, the automatic analysis of ancient Chinese is relatively struggling, making it difficult to meet the actual needs of Chinese historical studies and research. To promote the development of ancient Chinese resources and automatic analysis research, the Workshop on Language Technologies for Historical and Ancient languages (LT4HALA) of the International Language Resources and Evaluation Conference (LREC2022) aims at the task of word segmentation and part-of-speech (POS) tagging in Pre-Qin Chinese. The evaluation attempts to promote cooperation among scholars in related fields of ancient Chinese. This paper mainly conducts a series of research on word segmentation and part-of-speech tagging in ancient Chinese for this evaluation task.

Nowadays, research of the modern Chinese word segmentation and POS tagging tasks have achieved remarkable performance. Although the ancient Chinese word segmentation and POS tagging tasks are defined in the same way as the modern Chinese, they face many grammatical, lexical, and syntactic differences. In the fact of these differences, general modern Chinese word segmentation and POS tagging tools cannot accurately and effectively label ancient Chinese texts. In addition, previous methods proposed for modern Chinese have insufficient generalization ability for ancient Chinese due to the lack of annotated corpus. Fortunately, with the rapid development of deep learning technology, especially the emergence of pre-trained language models (PLMs) based on massive texts, the performance of deep learning models on many natural language processing (NLP) tasks in the ancient Chinese field has been greatly improved. Therefore, this paper jointly regards the ancient Chinese word segmentation and POS tasks as a joint sequence tagging task. Based on the PLM called SikuRoBERTa[1] provided by LREC2022, we add a layer of Conditional Random Field (CRF) to obtain more accurate label classification results. Besides, considering the limited amount of the provided ancient Chinese training data and the imbalance between various labels in POS tagging, we also employ various data augmentation techniques to improve the performance, including continual pre-training (Gururangan et al., 2020), adversarial training (Miyato et al., 2017), and ensemble learning. Extensive experiments conducted on the given

[*] Corresponding author.
[†] Equal contribution.
[1] https://huggingface.co/SIKU-BERT/sikuroberta

dataset demonstrate that our proposed models achieve comparable results, which reach the best F1-score of 0.9568 and 0.9114 on the online test set respectively.

2. Related Work

The task of word segmentation and POS tagging in the ancient Chinese field is not rare in NLP. Previous works on these tasks are mainly divided into the following two paradigms: (1) the step-by-step paradigm that firstly conducts word segmentation and then performs POS tagging and (2) the joint paradigm that deals with word segmentation and POS tagging at the same time. Unlike English sentences in which words are separated by spaces, Chinese sentences lack delimiter between words. Therefore, word segmentation is a fundamental step for the downstream tasks of Chinese NLP. However, when directly applying the modern Chinese word segmentation methods to ancient Chinese, it is hard to obtain an ideal effect due to the particularity of ancient texts. Hence, a more suitable word segmentation method must be proposed for ancient Chinese. For example, Gao and Zhao (2021) used a new word discovery method combining rules and statistics to discover new words from a large amount of classical literature and build an ancient Chinese word segmentation dictionary. Then the built dictionary is leveraged to segment the ancient texts. Traditional machine learning methods such as CRF combined with feature templates and professional dictionaries are employed to automatically segment ancient Chinese (Yang et al., 2017; Wang and Li, 2017). With the rapid progress of neural network (NN) technology, the Long Short-Term Memory (LSTM) and BERT models are also widely applied to the ancient Chinese word segmentation task (Gao, 2020; Gao, 2021). As for the ancient Chinese POS tagging task, researchers mainly employed rule-based methods (Liu and Dan, 2014) and traditional machine learning methods such as Hidden Markov Model (HMM) (Yang and Hu, 2020; Liang et al., 2002) and CRF model (Chiu et al., 2015).

The correctness of the POS tagging task somehow depends on the performance of word segmentation. However, the step-by-step paradigm would introduce multi-level diffusion of errors. Therefore, the joint paradigm of word segmentation and POS tagging tasks tend to bring more ideal results. Due to the particularity of ancient Chinese structure and semantics, expert knowledge would greatly affect the results of word segmentation and POS tagging. Hence, the method of leveraging rules and dictionaries is still commonly used in ancient texts (Li and Wei, 2013; Xing and Zhu, 2021). In addition, the performance of machine learning methods such as the maximum interval Markov network model (M3N) and CRF have been significantly improved (Qiao and Sun, 2010; Shi et al., 2010). In recent years, neural network models have been widely employed in ancient Chinese word segmentation and POS tagging tasks. Through integrating contextual and lexical information, the performance of POS tagging has been effectively improved (Cheng et al., 2020; Cui et al., 2020). In particular, Zhang et al. (2021) proposed a POS tagging model for ancient books based on the pre-trained language model BERT, which recently achieved the state-of-the-art performance for ancient Chinese POS tagging.

PLMs have become increasingly important in NLP. Extensive research has shown that PTMs trained on large corpora can learn general language representations, which can effectively improve the performance of downstream NLP tasks and avoid training new models from scratch. Undoubtedly, a suitable language model can greatly improve the model performance. Chinese is a language with unique features in syntax, vocabulary, and phonetics. Therefore, the Chinese PTMs should be in line with their unique characteristics. At present, scholars have proposed several PTMs for Chinese, including ERNIE (Li et al., 2019), CPM (Zhang et al., 2020), pre-trained Chinese language model using a whole-word masking strategy (Cui et al., 2021), and the fusion of glyph and pinyin information to Chinese BERT model (Sun et al., 2021).

3. Method

In this paper, we jointly model the ancient Chinese word segmentation and POS tagging tasks as a sequence labeling task. We adapt BERT-CRF as our base model and introduce four training methods to enhance the model performance, namely adversarial training (AT), continual pre-training, data augmentation[2] (DA), and ensemble learning.

3.1 Base Model

Our base model consists of two modules: the BERT encoder and the CRF output layer.

BERT. BERT is a transformer-based (Vaswani et al., 2017) pre-trained language model (PLM) that is designed to pre-train on a large unsupervised dataset to learn deep bidirectional representations. It consists of two subtasks, namely Mask Language Model (MLM) and Next Sentence Prediction (NSP). Being a variant of BERT, RoBERTa (Liu et al., 2019) aims to make full use of BERT architecture and training methods. There are three improvements in RoBERTa compared with BERT: (1) More training data; (2) Abondance of NSP task; and (3) Dynamic word masking.

We leverage the provided SikuRoBERTa to extract the representation for each token. After that, we leverage a softmax layer to produce the label scores for the tokens.

$$\boldsymbol{H} = SikuRoBERTa(\boldsymbol{X}) \quad (1)$$

$$\boldsymbol{P} = Softmax(W(\boldsymbol{H}) + b) \quad (2)$$

where W and b are parameters of the fully connected layer.

CRF. In the sequence labeling tasks, PLMs are hard to handle the dependency relationship between neighboring labels. In contrast, CRF can obtain an optimal prediction sequence by the relationship of neighboring labels, which can compensate for the shortcomings of PLMs. Thus, we further add a CRF layer to output the optimal label sequence $\boldsymbol{Y}^*$ for the input sequence.

$$\boldsymbol{Y}^* = CRF(\boldsymbol{P}) \quad (3)$$

3.2 Training Methods

Adversarial Training. AT is a training method that introduces adversarial perturbations to the original input to

[2] https://github.com/425776024/nlpcda

regularize the parameters and improve the robustness and generalization of the model. In this paper, we extend the fast gradient method (FGM) (which is originally proposed for text classification (Miyato et al., 2017)) to the ancient Chinese word segmentation and POS tagging tasks by adding the adversarial perturbations to the input token embedding of PLMs. The perturbations are calculated as follows:

$$r_{adv} = -\epsilon \frac{g}{||g||_2} \tag{4}$$

Where $\boldsymbol{g} = \nabla_x L(\theta, X, Y)$ is the model gradient.

Continual Pre-training. As stated in (Gururangan et al., 2020), it is helpful to tailor a PLM to the domain of a target task which can effectively enhance the performance of the target task. Therefore, we use the unsupervised data of the training set to continually pre-train the SikuRoBERTa to adapt the PLM to the *Zuo Zhuan* (the target domain of the given tasks) domain.

Data Augmentation. DA is a method of increasing the training data by adding small changes to the existing training data or creating new synthetic data from them. It can greatly alleviate the scenarios of insufficient data in deep learning. Given that the target task is typically a low-resource task with limited training data, we use a simple data augmentation approach with **identical label replacement** on the given training set. For example, given a sample of "*春秋左傳隱公*" , we replace "*隱公*" with the identically labeled word "*惠公*" to produce a new sample "*春秋左傳惠公*" . Through this method, we double the amount of training data.

Ensemble Learning. To further improve the generalization capability of the model, we use an ensemble learning approach to further fuse the results of multiple models. Specifically, using a voting mechanism, we vote on the results of multiple predictions for each word on a word-by-word basis. Then the label with the most votes is leveraged as the final output.

4. Experiment

4.1 Dataset

In this paper, the ancient Chinese annotation dataset from Zuozhuan (Li et al., 2013) provided by LT4HALA is used as the training set. After automatically segmented and tagged, the training set is then manually corrected by ancient Chinese experts. Finally, Chinese words are separated by spaces, and each token is tagged as "word/tag" format like "*隱公/nr*". We perform a statistical analysis of the POS labels in the training set, and the results are shown in Figure 1. Results show that punctuation marks (w), verbs (v), and nouns (n) appear most frequently, of which punctuation marks appear 42,315 times. However, rare categories such as rn, nn, nsr, and rr appear very few times, resulting in an imbalance between POS categories on the training set, which has also become an important improvement direction we consider when optimizing the model. In the testing phase, we used two test sets, Test A and Test B. Test A is extracted from the same book named *Zuozhuan* as the training set, which is used to evaluate the model's ability to recognize the same source but non-overlapping data. Test B is extracted from other ancient Chinese books to evaluate the generalization ability of the model in a similar ancient Chinese corpus. The dataset statistics are shown in Table 1.

Datasets	Word Tokens	Char Tokens
Train	166,142	194,995
Test A	28,131	33,298
Test B	Around 40,000	Around 50,000

Table 1: The dataset statistics.

4.2 Experimental Settings

Our models are all implemented based on the PyTorch[3] framework. As mentioned above, PTMs are proven to achieve considerable performance on ancient Chinese word segmentation and POS tagging tasks. Therefore, we first conduct a series of base experiments on different Chinese PTMs, including guwenbert-base[4], chinese-roberta-wwm-ext[5], roberta-classical-chinese-base-char[6], and the provided SikuRoBERTa. Through experimental comparison, we choose SikuRoBERTa to optimize in subsequent work because SikuRoBERTa performs better than other PTMs.

Next, we carry out extensive experiments on optimizing model parameters to fine-tune the model. Specifically, we set epoch to {2, 3, 4}; learning rate as {1e-4, 2e-5, 5e-5, 10e-6}; loss function as {CrossEntropy, Focal Loss}; batch size as {16, 24}. Since there exist several long sentences, we always set the maximum sequence length as 128 in training and 512 in inference.

As for the evaluation, we use the average F1-score over five cross-validation folds on dev data as the offline test set to represent the performance during our training phase and the official F1-score for the final online evaluation on both word segmentation and POS tagging tasks.

4.3 Results and Analysis

The results of our models on both word segmentation and part-of-speech tagging task are illustrated in Table 2. As shown in the table, all models achieve better results on word segmentation than POS tagging, among which CP_SikuRoBERTa_CRF_ADV achieves the best F1-score on Test A of 0.9568. As for the POS tagging task, the introduced continual pre-training, the added CRF layer, and adversarial training strategies all yield better performance when compared to the baseline model SikuRoBERTa-softmax. Among them, CP_SikuRoBERTa_CRF_ADV model achieves the best F1-score on Test A of 0.9114. In addition, the models perform better on Test A than Test B since Test B is extracted from a different ancient Chinese book than the training set, but also achieves an F1-score of 0.8699.

[3] https://pytorch.org/

[4] https://huggingface.co/ethanyt/guwenbert-base

[5] https://huggingface.co/hfl/chinese-roberta-wwm-ext

[6] https://huggingface.co/KoichiYasuoka/roberta-classical-chinese-base-char

Num.	Model	Evaluation Type	Test Set	Word Segmentation	Pos Tagging
1	guwenbert-base	Offline	test	0.9433	0.8595
2	chinese-roberta-wwm-ext		test	0.9468	0.8773
3	roberta-classical-chinese-base-char		test	0.9519	0.8823
4	CP_SikuRoBERTa_CRF		test	0.9520	0.8845
5	DA_SikuRoBERTa-softmax		test	0.9508	0.8840
6	SikuRoBERTa-softmax	Online	Test A	0.9363	0.8932
			Test B	0.9308	0.8667
7	CP_SikuRoBERTa-softmax		Test A	0.9365	0.8937
			Test B	0.9248	0.8588
8	CP_SikuRoBERTa_CRF_ADV		**Test A**	**0.9568**	**0.9114**
			Test B	**0.9364**	**0.8699**
9	ensemble（6+7+8）		Test A	0.9384	0.8964
			Test B	0.9301	0.8676

Table 2: The results of our models on the EvaHan shared task. In this table, CP indicates continual pre-training, DA indicates data augmentation, and ADV indicates adversarial training.

Surprisingly, it seems that the DA strategy can not improve the model performance. One possible reason might be that the augmentation method is somehow simple and often unavoidably generates many meaningless duplicate data. Considering that the training data size is small and thus this simple DA strategy does not work well.

In addition, as demonstrated in the table, we perform an ensemble learning strategy on the three models with the best performance. Although the final result is better than the single SikuRoBERTa-softmax and CP_SikuRoBERTa softmax models but is worse than the single CP_SikuRoBERTa_CRF_ADV model. We speculate the possible reason is that the difference between the three models is not obvious, and they both cannot handle well with some rare categories, resulting in unsatisfactory results in the final integration.

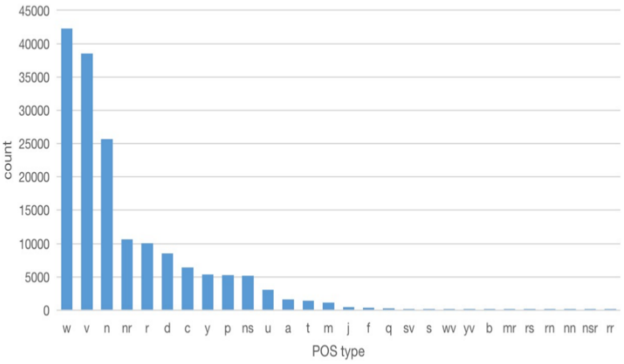

Figure 1: The frequency distribution of part-of-speech tags in the training data.

5. Conclusion

We present our results for the EvaHan 2022 shared task at International Language Resources and Evaluation Conference (LREC2022). We model the ancient Chinese word segmentation and POS tagging tasks as a joint sequence tagging problem and employ augmentation methods such as continuous pre-training, adversarial training, and ensemble learning after analyzing the training set. Overall, our model performs well on both word segmentation and POS tagging on the official test set, with the best F1-scores of 0.9568 and 0.9114 respectively. However, the ensemble learning strategy seems not to improve the performance as we expected. Additionally, the methods we tried are all implemented under a closed modality that only trained on the official training set, and do not consider the importance of other external data resources that might bring improvement. In future work, on the one hand, we plan to explore more efficient ensembles to improve the model performance. On the other hand, we will investigate the use of external ancient Chinese resources to further improve the model performance.

6. Acknowledgment

The work is supported by the Humanity and Social Science Youth Foundation of Ministry of Education of China (18YJCZH024) and Guangdong University of Foreign Studies (299-X5219112, 299-X5218168).

7. References

Cheng, N., Li, B., Ge, S. J., Hao, X. Y., and Feng, M. Y. (2010). A joint model of automatic sentence segmentation and lexical analysis for ancient Chinese based on BiLSTM-CRF model. *Journal of Chinese Information Processing*, (4):1-9.

Chiu, T. S., Lu, Q., Xu, J., Xiong, D., and Lo, F. (2015). PoS tagging for classical Chinese text. *CLSW.*

Cui, D. D., Liu, X. L., Chen, R. Y., Liu, X. H., Li, Z., and Qi, L. (2020). Named entity recognition in field of ancient Chinese based on Lattice LSTM. *Computer Science*, 47(S02):18-22.

Cui, Y. M., Che, W. X., Liu, T., Qin, B., Yang, Z. Q., Wang, S. J., and Hu, G. P. (2021). Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504-3514.

Gao, J. C. and Zhao, Q. C. (2021). Study on word segmentation method of classical literature based on new word discovery. *Computer Technology and Development*, 31(09):178-181+207.

Gao, Y. (2020). Ancient Chinese word segmentation system based on long and short time neural network. *Automation and Instrumentation*, (2):128-131.

Gao, Y. (2021). Research on automatic word segmentation method of ancient Chinese based on BERT prediction training model. *Electronic Design Engineering*, (22):28-32.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* pages 8342–8360, Online. Association for Computational Linguistics.

Huang, L., Peng, Y. N., Wang, H., and Wu, Z. Y. (2002). Statistical part-of-speech tagging for classical Chinese. In *Text, Speech and Dialogue, 5th International Conference, TSD 2002, Brno, Czech Republic September 9-12, 2002, Proceedings. DBLP, 2002*. pp.115-122.

Li, B., Feng, M. X., and Chen, X. H. (2013). Corpus Based Lexical Statistics of Pre-Qin Chinese. *Lecture Notes in Computer Science*, 7717:145-153.

Li, X. Y., Meng, Y. X., Sun, X. F., Han, Q. H., Yuan, A., and Li, J. W. (2019). Is word segmentation necessary for deep learning of Chinese representations? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3242–3252.

Liu, Y. and Long, D. (2014). A rule-based method for identifying patterns in old Chinese sentences. CLSW.

Liu, Y. H., Ott, M., Goyal, N., Du, J. F., Joshi, M., Chen, D. Q., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Miyato, T., Dai, A. M., and Goodfellow, I. J. (2017). Adversarial training methods for semi-supervised text classification. In *Proceedings of the 2017 International Conference on Learning Representations (ICLR)*, pp. 1-11.

Qiao, W. and Sun, M. S. (2010). Joint Chinese word segmentation and named entity recognition based on max-margin Markov networks. Journal of Tsinghua University(Science and Technology), 50(5):758-762,767.

Qin, L. and Wei, W. (2013). Research on the system of jointing Chinese word segmentation with part-of-speech tagging. *2013 Sixth International Symposium on Computational Intelligence and Design*, 1:387-390.

Shi, M., Li, B., and Chen, X. H. (2010). CRF based research on a unified approach to word segmentation and POS tagging for Pre-Qin Chinese. *Journal of Chinese Information Processing*, 24(2):39-45.

Sun, Z. J., Li, X. Y., Sun, X. F., Meng, Y. X., Ao, X., He, Q., Wu, F., and Li, J. W. (2021). ChineseBERT: Chinese pre-training enhanced by glyph and pinyin information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2065–2075, Online. Association for Computational Linguistics.

Wang, X. Y. and Li, B. (2017). Automatically segmenting middle ancient Chinese words with CRFs. *Data Analysis and Knowledge Discovery*, 1(5): 62-70.

Yang, S. C., Ji, Y., and Zhao, L. P. (2017). Study of ancient Chinese word segmentation based on Conditional Random Field. *Computer Knowledge and Technology*, (22):183-184.

Yang, X. S. and Hu, L. S. (2020). Part-of-speech tagging of classical Chinese based on Hidden Markovian Model. *Microcomputer Applications*, 36(5):130-133.

Xing, F. G. and Zhu, T. S. (2021). Large-scale online corpus based classical integrated Chinese dictionary construction and word segmentation. *Journal of Chinese Information Processing*, 35(7):41-46.

Zhang, Q., Jiang, C., Ji, Y.S., Feng, M. X., Li, B., Xu, C., and Liu, L. (2021). Unified model for word segmentation and POS tagging of multi-domain Pre-Qin literature. *Data Analysis and Knowledge Discovery*, 5(3):2-11.

Zhang, Z. Y., Han, X., Zhou, H., Ke, P., Gu, Y. X., Ye, D. M., Qin, Y. J., Su, Y. S., Ji, H. Z., Guan, J., Qi, F. C., Wang, X. Z., Zheng, Y. A., Zeng, G. Y., Cao, H. Q., Chen, S., Li, D. X., Sun, Z. B., Liu, Z. Y., Huang, M. L., Han, W. T., Tang, J., Li, J. Z., Zhu, X. Y., and Sun, M. S. (2020). CPM: A large-scale generative Chinese pre-trained language model. *ArXiv*, abs/2012.00413.