Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 174–177
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022

A Joint Framework for Ancient Chinese WS and POS Tagging based on Adversarial Ensemble Learning

Shuxun Yang

School of Computer Science, Beijing Institute of Technology, Beijing, China
sheryl_xun@163.com

Abstract

Ancient Chinese word segmentation and part-of-speech tagging tasks are crucial to facilitate the study of ancient Chinese and the dissemination of traditional Chinese culture. Current methods face problems such as lack of large-scale labeled data, individual task error propagation, and lack of robustness and generalization of models. Therefore, we propose a joint framework for ancient Chinese WS and POS tagging based on adversarial ensemble learning, called AENet. On the basis of pre-training and fine-tuning, AENet uses a joint tagging approach of WS and POS tagging and treats it as a joint sequence tagging task. Meanwhile, AENet incorporates adversarial training and ensemble learning, which effectively improves the model recognition efficiency while enhancing the robustness and generalization of the model. Our experiments demonstrate that AENet improves the F1 score of word segmentation by 4.48% and the score of part-of-speech tagging by 2.29% on test dataset compared with the baseline, which shows high performance and strong generalization.

Keywords: Adversarial Ensemble Learning, Word Segmentation, POS Tagging

1. Introduction

Recently, researchers have gradually paid more attention to traditional culture, and the understanding and study of ancient Chinese is an important parts. However, there are many obstacles in understanding ancient Chinese due to the features of the separation of language and text, archaic and incomprehensible, and unclear segmentation. In order to better help researchers understand ancient Chinese and promote the inheritance of Chinese traditional culture, applying some basic tasks of natural language processing (NLP), such as word segmentation (WS), part-of-speech (POS) tagging, and named entity recognition (NER), to ancient Chinese has become an urgent need.

Chinese word segmentation, refers to the partitioning of a sequence of consecutive words in units of words into word-based sequences by word segmentation algorithms with the help of computer technology. Part-of-speech tagging refers to tagging the words in a sentence by part-of-speech tagging algorithms, that is, predicting the lexicality of words. These two tasks are the basis of many downstream tasks of natural language processing and play an indispensable role in various fields.

In fact, both the WS and POS tagging tasks can generally be regarded as sequence labeling tasks. Defining a suitable labeling scheme provides ideas to solve these problems. Due to the large differences between ancient Chinese and modern texts, the difficulty of understanding and the lack of obvious segmentation symbols, early ancient Chinese WS and POS tagging tasks would often be solved by taking a manual construction approach. These methods tend to have a high accuracy rate, with unacceptable cost. After that, methods based on lexical, dictionaries, and manual rules emerged. Researchers find strings that match those rules with the help of priority rules constructed manually by experts in various fields. However, these methods rely on the construction of dictionaries and knowledge bases, and system constructed tend to be less portable and scalable, and probably require experts in specific domain to spend a lot of time on construction and maintenance.

With the development of computer technology, the demand for automatic WS and POS tagging of ancient Chinese has increased, and algorithms based on machine learning and deep learning have emerged. Conditional Random Fields (CRF), Support Vector machines (SVM), Hidden Markov Models (HMM), Maximum Entropy Models (MEM), Long Short-Term Memory Networks (LSTM), Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN) and so on are widely used in WS and POS tagging task of ancient Chinese. However, supervised learning methods above usually require large-scale labeled datasets, and the field of ancient Chinese often faces the problem of sparse labeled data. Therefore, pre-trained language models (PLM) with fine-tuning have come into the forefront of researchers' attention. This approach essentially uses transfer learning to train a word vector model with rich semantic information using a large amount of unlabeled text, and then fine-tune it using labeled data, which can well solve the problem of lacking high-quality, large-scale labeled data in a specific domain.

However, WS and POS tagging models in modern standard Chinese often do not work well for ancient Chinese, and the trained models are often sensitive to noisy data and do not have good portability and transferability. Adversarial training (AT) and ensemble learning (EL) can help us solve these problems well. Adversarial training is an important way to enhance the robustness of neural networks. The essential idea of AT is adding some small but potentially misclassifying perturbations to the samples during training process will make the model adapt to such changes and thus be robust to the adversarial samples. Ensemble learning, on the other hand, as a common approach for supervised machine learning tasks, aims to improve the prediction results by the integration of multiple learning algorithms. Combing adversarial training with ensemble learning can enhance the portability and robustness of the model while improving the accuracy of ancient Chinese WS and POS tagging tasks.

In summary, we propose a joint framework based on adversarial ensemble learning for ancient Chinese WS and POS tagging tasks, called AENet, to address the problems of lack of large-scale annotation data, low model portability and robustness for joint tasks of ancient Chinese WS and POS tagging. The main innovations of this paper are as follows.

- We propose a joint framework for ancient Chinese WS and POS tagging to reduce the noise caused by individual task training process and improve recognition efficiency of the model, with the idea of pre-training and fine-tuning.
- We incorporate the ideas of adversarial training and ensemble learning into the joint framework to improve the robustness and generalization of our model effectively.
- Compared with baseline, the proposed framework achieves better performance on two ancient Chinese datasets provided.

2. Related Work

With the deepening on ancient Chinese mining research, researchers are in full swing on the study of ancient Chinese WS and POS tagging tasks. For example, Yu et al. (2020) proposed an automatic WS model for ancient Chinese based on a nonparametric Bayesian model and deep learning. This method adopts an unsupervised multi-stage iterative training, aiming to mine valuable ancient Chinese WS models by jointly using Bayesian model and BERT, and training them repeatedly in large-scale unlabeled data. Cheng et al. (2020) designed an ancient Chinese WS and POS tagging model based on BiLSTM-CRF model, and by designing appropriate WS and POS labels, these two tasks were fused, which is similar to the method of task fusion in this paper. Stoeckel et al. (2020) proposed an ensemble classifier, namely LSTMVote, for the POS tagging task of Latin languages, which integrates multiple pre-trained classifiers to obtain the optimal model.

To solve the problem of lack of ancient Chinese annotated corpus, pre-trained language models have been introduced to the study. Based on the ancient literature corpus of Daizhige [1], GuwenBERT [2] model was proposed. This method combines the weight of modern Chinese RoBERTa model and a large number of ancient Chinese corpus on the basis of the continuation training technique, and transfers some linguistic features of modern Chinese to ancient Chinese, which substantially improves the performance of the model. After that, Wang et al. (2021) constructed SikuBERT and SikuRoBERTa pre-trained language models for ancient Chinese intelligent processing tasks based on the BERT, using the calibrated high-quality full-text corpus of *Siku Quanshu* as an unsupervised training set, which provided support for researchers in ancient Chinese.

Numerous studies have proved that adversarial training can effectively improve the robustness and generalization of language models. FGSM and FGM adversarial training methods (Goodfellow et al., 2014; Miyato et al., 2017) were proposed, the core idea of which is to let the direction of perturbation follow the direction of gradient boosting. In these methods, authors assume that the loss function is linear or locally linear, and therefore the direction of gradient boosting is the optimal direction. The difference between FGSM and FGM is the normalization method, with FGSM taking max normalization of the gradient through the sign function and FGM using L2 normalization. In order to solve the linear assumption problem in FGSM and FGM, Projected Gradient Descent method (PGD) (Madry et al., 2017) was proposed, which can be used to solve the internal maximum problem. The core idea of PGD is to reach the optimum by multiple iterations and each iteration will project the perturbation to a specified range. However, this method can only utilize the gradient of the parameters and the gradient of the input alone. In order to utilize two gradients simultaneously and efficiently, FreeLB (Zhu et al., 2019) was proposed, which makes use of the gradient accumulated from multiple iterations to make updates and estimate the gradient more accurately.

Meanwhile, as an effective way of supervised learning, ensemble learning can obtain better prediction performance than using any individual learning algorithm alone by integrating multiple learning algorithms. At present, ensemble learning algorithms are mainly classified into three categories: Bagging, Boosting and Stacking, which correspond to parallel training, serial training and hierarchical training, respectively. With the help of the idea of ensemble learning, Izmailov et al. (2018) proposed a stochastic weight averaging (SWA) algorithm, whose core idea is that the average of multiple weights in the training process of a single model is closer to the optimal solution. A lot of practices have proved that SWA is superior to other optimization algorithms, such as SGD.

3. Model

In this section, we first introduce the task definition, and then present the overall framework of the joint model for ancient Chinese WS and POS tagging tasks. After that, we detail how to jointly use adversarial training and ensemble learning to improve model performance.

3.1 Task Definition

Given an input sentence of ancient Chinese with *n* tokens $X=\{x_1,x_2\cdots x_n\}$, the target sentence can be $Y=\{y_1,y_2\cdots y_n\}$, where $y_i='ws_p-pos_q'$, for example, $y_1='B\text{-}NR'$. In the formula above, $ws_p=\{B,I,E,S\}$. *B* means the current token is the beginning of a multi-token word, *I* means the current token is in the middle of a multi-token word, *E* means the current token is the end of a multi-token word, and *S* means the current token is an single word. Through this tagging method, the task of WS for ancient Chinese can be solved automatically. And then, $pos_q=\{A,C,D,J,M,N,NR,NS\cdots\}$, which refers to common parts of speech in texts. The task in this paper can be defined in the form of Equation 1, that is, given a sequence *X*, find the optimal sequence *Y* that maximizes the probability of $p(Y\mid X)$. According to the above tagging methods, the joint task of WS and POS tagging of ancient Chinese can be realized easily, thus reducing the noise impact and error propagation that may be brought by separate task training.

$$Y^*=\arg\max p(Y\mid X) \tag{1}$$

3.2 Model Framework

The overall joint framework for ancient Chinese WS and POS tagging based on adversarial ensemble learning, that is, AENet, is shown in Figure 1. The overall framework of AENet is carried out with the idea of pre-training and fine-

[1] http://www.daizhige.org/

[2] https://github.com/ethan-yt/guwenbert

tuning. Namely, given an ancient Chinese sequence, it is cut into token sequences firstly. Then, the token sequence is input into the pre-trained language model for fine-tuning, and word embeddings with rich semantic information can be obtained. The final predicted label sequences are obtained by feeding word embeddings into the CRF layer. The specific process is shown in Equation 2.

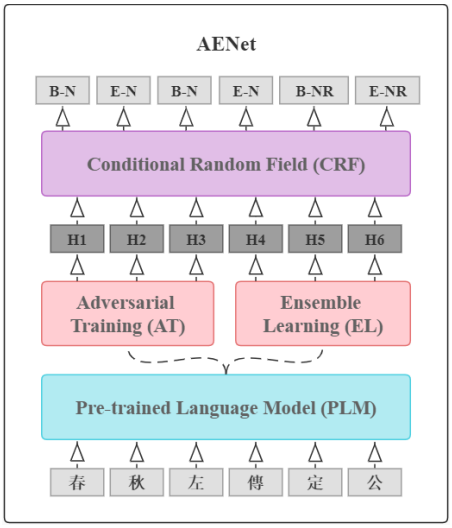

Figure 1: Framework of AENet

Throughout the entire model training process, AENet will be optimized according to the adversarial training and ensemble learning, thereby enhancing the robustness and generalization of the model. See Section 3.3 for details. The loss function of AENet is the log-likelihood function, as shown in Equation 3.

$$\begin{aligned} Embedding &= PLM(X) \\ Y &= CRF(Embedding) \end{aligned} \tag{2}$$

$$Loss = -\log P(Y \mid X) \tag{3}$$

3.3 Adversarial Ensemble Learning

The idea of adversarial training is to add some small but potentially misclassifying perturbations to the samples during the training process of the model, making the model adapt to such changes and thus increasing the robustness and the transferability of the model. The process of adversarial training is shown in Equation 4, where δ represents the perturbation, ε is a parameter set in advance to constrain the range of the perturbation and w is the model weight with parameters θ. Equation 4 means the whole process of model optimization, that is, finding the perturbation that maximizes the loss function and training the neural network model to minimize its loss on the training data after superimposing the perturbation.

$$\begin{aligned} \delta &= \arg\max_{\|\delta\| \le \varepsilon} Loss\left(f_\theta(X+\delta), Y\right) \\ X &= X + \delta \\ w(\theta) &= \arg\min Loss(X, Y) \end{aligned} \tag{4}$$

In this paper, we select FGM adversarial training method (Miyato et al., 2017), and the perturbation parameters are calculated as shown in Equation 5, where g represents the gradient of the loss function. During each training of the model, we calculate the perturbation and add it to the training samples, so that the trained model is sufficient to cope with the perturbation and increase the robustness.

$$\begin{aligned} \delta &= \varepsilon \cdot \left(g / \|g\|_2\right) \\ g &= \nabla_X \left(Loss\left(f_\theta(X), Y\right)\right) \end{aligned} \tag{5}$$

Meanwhile, during the overall training process of AENet, we optimize the model weights with the help of ensemble learning ideas and SWA model (Izmailov et al., 2018). The final weights of the model are calculated by Equation 6, where n, m are the parameters set in advance.

$$\bar{w}(\theta) = 1/(n-m+1)\sum_{i=m}^{n} w_i(\theta) \tag{6}$$

After incorporating adversarial training and ensemble learning into the joint framework, the whole model of ancient Chinese WS and POS tagging based on adversarial ensemble learning, namely AENet, is constructed in this paper.

4. Experiment

4.1 Experimental Setup

The experiments in this paper are conducted on a server with Ubuntu 20.04 Linux and eight 1080Ti GPUs. The code is written in Python 3.8.5 environment using PyTorch. We carry out these experiments for the EvaHan 2022 competition. This contest is divided into two modalities: closed and open. In the closed modality, only the provided training dataset and the SikuRoBERTa pretrained model are allowed to be used. In this paper, the closed modality is selected for the experiments. Therefore, the SikuRoBERTa is used for the pre-trained language model in the AENet model framework. The parameter ε in the adversarial training is set to 1, and n in the ensemble learning is set to 5 while m is set to 1. Precision, recall, and F1 score metrics are used to evaluate the results of ancient Chinese WS and POS tagging, respectively.

4.2 Dataset Description

The training data and test data involved in the experimental part of this paper are provided by the organizer of EvaHan 2022 competition. The training data is selected from *Zuozhuan*, an ancient Chinese work believed to date from the Warring States Period, which contains punctuation and ancient Chinese texts after WS and POS tagging, and is presented in the form of utf-8 plain text files. The training data has a total of 166142 word tokens and 194995 char tokens.

The test dataset is divided into test A and B. Test A is still extracted from *Zuozhuan*, which does not overlap with the training data, mainly to observe the performance of the model in the text data of the same book. Test A mainly consists of 28131 word tokens and 33298 char tokens. Test B dataset is extracted from other books, mainly to observe the performance of the model in similar text data. Its size is similar to the test A dataset.

4.3 Experimental Results

In this section, CRF and SikuRoBERTa + BiLSTM + CRF models are selected as baselines, to compare with AENet model we proposed. The running results of CRF model are provided by EvaHan 2022 organizers. The experimental results for test A dataset are shown in Table 1, and the experimental results for test B dataset are shown in Table 2.

Metric(%)	Precision	Recall	F1 score
CRF (WS)	90.64	92.08	91.35
CRF (POS)	89.06	89.54	89.30
PLM+BiLSTM+CRF (WS)	95.15	96.07	95.61
PLM+BiLSTM+CRF (POS)	90.69	91.56	91.12
AENet (WS)	**95.18**	**96.49**	**95.83**
AENet (POS)	**90.96**	**92.22**	**91.59**

Table 1: Results for test A

Metric(%)	Precision	Recall	F1 score
PLM+BiLSTM+CRF (WS)	93.49	90.39	91.91
PLM+BiLSTM+CRF (POS)	87.02	84.14	85.56
AENet (WS)	**94.48**	**91.70**	**93.07**
AENet (POS)	**88.40**	**85.80**	**87.08**

Table 2: Results for test B

The experimental results show that the use of the model framework of pre-training and fine-tuning substantially improved the performance of the model. In the test A dataset, compared with the baseline CRF model, AENet improves the F1 score of WS by 4.48% and the score of POS tagging by 2.29%.

In addition, we find that although the WS task of the AENet model is 0.22% higher than the SikuRoBERTa + BiLSTM + CRF model and the POS tagging task improves 0.47% in the test A, the WS task of the AENet model is 1.16% higher than the SikuRoBERTa + BiLSTM + CRF model in the test B and the POS tagging task improves by 1.52%. This is sufficient to demonstrate that the robustness and generalization of our AENet model are substantially improved by introducing adversarial ensemble learning.

4.4 Ablation Study

This section focuses on the ablation analysis of the AENet model and observes the degree of influence of adversarial training and ensemble learning on the robustness and generalization of the model. Therefore, we compare the model using only adversarial training, that is, AENet_{AT} and only ensemble learning, that is, AENet_{EL} with the original AENet model, and the experimental results for test B dataset are shown in Table 3.

Metric(%)	Precision	Recall	F1 score
PLM+BiLSTM+CRF (WS)	93.49	90.39	91.91
PLM+BiLSTM+CRF (POS)	87.02	84.14	85.56
AENet_{AT} (WS)	93.93	90.66	92.26
AENet_{AT} (POS)	87.91	84.85	86.35
AENet_{EL} (WS)	94.39	91.58	92.96
AENet_{EL} (POS)	87.83	85.21	86.50
AENet (WS)	**94.48**	**91.70**	**93.07**
AENet (POS)	**88.40**	**85.80**	**87.08**

Table3: Ablation study results for test B

It is experimentally demonstrated that compared to baseline, both adversarial training and ensemble learning improve the performance of our model for WS and POS tagging in similar ancient Chinese texts, and AENet achieves the best performance by integrating AT and EL. There is no doubt that adversarial ensemble learning in AENet improves the robustness and generalization of the model.

5. Conclusion

We introduce a joint framework based on adversarial ensemble learning in this paper, namely AENet, for the task of ancient Chinese WS and POS tagging. On the basis of pre-training and fine-tuning, AENet treats WS and POS tagging as a joint sequence tagging task, and we design a joint tagging approach to reduce the error propagation and noise impact caused by individual task training. Then, AENet incorporates adversarial training and ensemble learning, which effectively enhances the robustness and generalization of the model we proposed while improving the recognition efficiency of the model. The experimental results demonstrate that AENet has better performance in handling the ancient Chinese WS and POS tagging tasks, compared with baselines.

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*, pp. 52-58.

Goodfellow, I.J., Shlens, J. and Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *Stat*, 1050, p.20.

Izmailov, P., Wilson, A.G., Podoprikhin, D., Vetrov, D., and Garipov, T. (2018). Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, pp. 876-885.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Miayto, T., Dai, A.M. and Goodfellow, I. (2016). Virtual Adversarial Training for Semi-Supervised Text Classification.

Stoeckel, M., Henlein, A., Hemati, W., and Mehler, A. (2020). Voting for POS tagging of Latin texts: Using the flair of FLAIR to better ensemble classifiers by example of Latin. In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*, pp. 130-135.

Wang, D., Liu, C., Zhu, Z., Liu, J., Hu, H., Shen, S., and Li, B. (2021). Construction and Application of Pre-training Model of "Siku Quanshu" Oriented to Digital Humanities. *Library Tribune*.

Yu, J., Wei, Y., Zhang, Y., and Yang, H. (2020). Word Segmentation for Ancient Chinese Texts Based on Nonparametric Bayesian Models and Deep Learning. *Journal of Chinese Information Processing*, 34(6): 1-8.

Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., and Liu, J. (2019). FreeLB: Enhanced Adversarial Training for Natural Language Understanding. In *International Conference on Learning Representations*.