Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 141–145
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022

Automatic Word Segmentation and Part-of-Speech Tagging of Ancient Chinese based on BERT Model

CHANG Yu[1], ZHU Peng[1], WANG Chaoping[2], WANG Chaofan[3]
1.School of Data Science and Application, Inner Mongolia University of Technology, Hohehot Municipality, Inner Mongolia, China
2. Institute of Sinology, Nanchang University, Nanchang, Jiangxi, China
3. College of Acupuncture & Tuina and Rehabilitation, Hunan University of Traditional Chinese Medicine, Changsha, Hunan, China
20211100482@imut.edu.cn, 20211800690@imut.edu.cn, 2273718186@qq.com, 2272592717@qq.com

Abstract

In recent years, new deep learning methods and pre-training language models have been emerging in the field of natural language processing (NLP). These methods and models can greatly improve the accuracy of automatic word segmentation and part-of-speech tagging in the field of ancient Chinese research. In these models, the BERT model has made amazing achievements in the top-level test of machine reading comprehension SQuAD-1.1. In addition, it also showed better results than other models in 11 different NLP tests. In this paper, *SIKU-RoBERTa* pre-training language model based on the high-quality full-text corpus of *SiKuQuanShu* have been adopted, and part corpus of *ZuoZhuan* that has been word segmented and part-of-speech tagged is used as training sets to build a deep network model based on BERT for word segmentation and POS tagging experiments. In addition, we also use other classical NLP network models for comparative experiments. The results show that using *SIKU-RoBERTa* pre-training language model, the overall prediction accuracy of word segmentation and part-of-speech tagging of this model can reach 93.87% and 88.97%, with excellent overall performance.

Keywords: natural language processing, deep learning, BERT model, automatic part-of-speech tagging

1. Introduction

At present, the automatic lexical analysis technology for modern Chinese (including automatic word segmentation, part of speech tagging, named entity recognition, etc.) has been basically mature. People try to use the existing modern Chinese analysis model to deal with ancient Chinese. However, due to the use of traditional characters in ancient Chinese, it first needs to spend a lot of resources and time to convert traditional characters into simplified characters. Secondly, there are significant differences between ancient Chinese and modern Chinese in font, vocabulary and grammar. Finally, more ancient Chinese texts lack sentence breaks and punctuation, which brings great difficulties to further data analysis, knowledge mining and the development of related intelligent applications.

The research on automatic word segmentation of ancient Chinese has also experienced three stages: rule-based system, statistics-based method and deep-learning-based network model.

Huang et al. designed the automatic word segmentation algorithm of agricultural ancient books through *N-ary* grammar and dictionary word segmentation technology. After testing, it has a good word recognition rate on 13 agricultural ancient books. Xu et al. proposed a rule-based word segmentation method for *ZuoZhuan*, and the F1 value of this method reached 89.46%.

Fang et al. proposed a word segmentation algorithm based on likelihood ratio statistical method, and realized the automatic word segmentation of tea classic through tree pruning algorithm. Chen et al. constructed an improved statistical model of ancient Chinese text based on Kalman filter. Compared with the baseline model, the accuracy of word segmentation in *ShiJi* and *Song History* increased by 30%.

Wang et al. determined the combined feature template through conditional random field model and statistical method, and finally obtained the part of speech automatic annotation algorithm model for Pre-Qin classics. The harmonic average value f of the model reaches 94.79%. Cheng et al. proposed an integrated annotation method of sentence segmentation and lexical analysis based on BiLSTM-CRF neural network model. The F1 value of word segmentation task and part-of-speech tagging task on the comprehensive test set of the model reached 85.73% and 72.65%.

SIKU-RoBERTa is a natural language pre-training model based on BERT model and trained with *SiKuQuanShu*. This experiment will use part of the *ZuoZhuan* as the training set, fine-tune on the basis of *SIKU-RoBERTa*, and complete the tasks of word segmentation and part-of-speech tagging. In addition, some classical natural language processing models will be used as comparative experiments.

2. Model Introduction

BERT model is a pre-training language model proposed by Google, which breaks through the limitation that text representation methods such as one-hot and word2vec can only generate a word vector for each word in the thesaurus, and solves the thorny problem of polysemy. In addition, based on the self-attention mechanism, BERT model can contain deeper context information, which plays a decisive role in the effect of natural language processing tasks. It is a milestone in the research of natural language processing. It has set a new record in 11 natural language processing tasks and has become the focus of current research.

The basic BERT model is composed of 12 layers of transformer encoder units, each layer has 12 Attention, and the hidden layer size H is 768, that is, the word vector dimension. Its structure is shown in Figure 1.

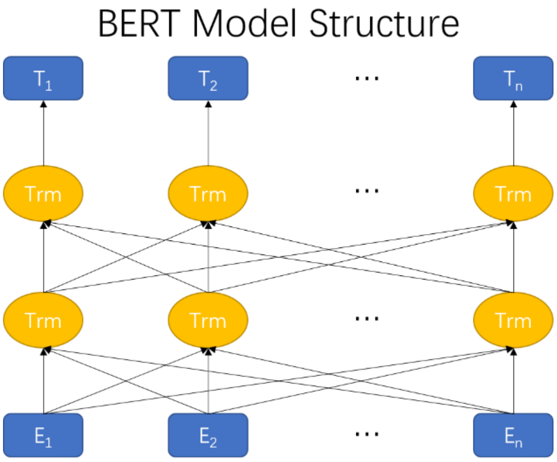

Figure 1: BERT model structure

The intermediate feature extraction adopts the structure of the encoder part of the transformer, but uses a new activation function GeLU (Gaussian error linear unit) instead of the original activation function ReLU of the transformer.

The embedding layer of BERT model is shown in Figure 2, which is obtained by the superposition of *Token Embedding*, *Segment Embedding* and *Position Embedding*. The *Segment Embedding* can be used for sentence classification tasks, such as judging whether the two sentences are semantically similar and whether the two sentences are context, etc.

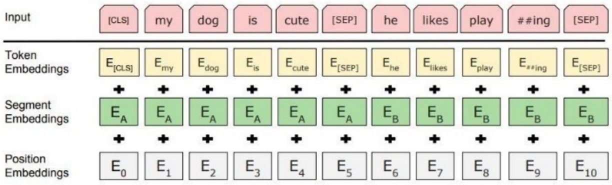
Input
[CLS]
my
dog
is
cute
[SEP]
he
likes
play
##ing
[SEP]
Token Embeddings
E[CLS] Emy Edog Eis Ecute E[SEP] Ehe Elikes Eplay E##ing E[SEP]
Segment Embeddings
EA EA EA EA EA EA EB EB EB EB EB
Position Embeddings
E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 E10

Figure 2: Embedding layer of BERT model

According to the output structure of the BERT model, inputting the output of the BERT model into a Full Connection Layer, each token of the input sentence can be labeled to complete the sequence labeling task, and then complete the tasks such as word segmentation, part-of-speech tagging, named entity recognition and so on. Its structure is shown in Figure 3.

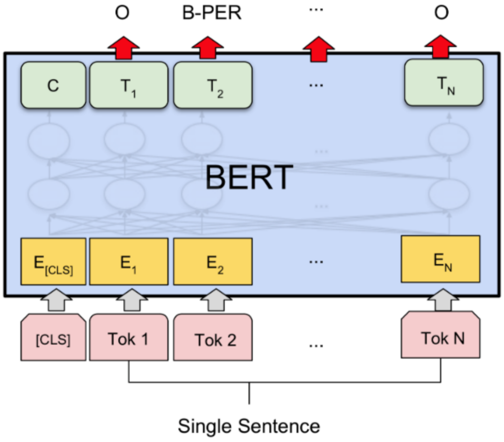

Figure 3: The structure of sequence annotation

The Data Set	Characters Quantity	Words Quantity	POS Distribution(Top 5)

Training Set	186282	157441	Verb(24.4%), Punctuation(21.3%)
			Noun(16.3%), Person(6.7%), Pronoun(6.4%)

Test Set A	33297	28131	Punctuation(26.1%), Verb(23.7%)
			Noun(12.9%), Person(7.2%), Pronoun(6.1%)

Test Set B	62969	53835	Verb(23.5%), Pronoun(21.7)
			Noun(13.8%), Person(7.9%), Location(6.2%)

Table 1: overview of experimental data set

3. Data Set Introduction

3.1 Datasets with Labels

The training set used in this experiment is from *ZuoZhuan(左 传)*, which has been segmentation and marked with part-of-speech. The sentence segmentation is realized by the end of the sentence in the corpus, such as period, question mark, exclamation mark and other symbols. The source language in the final training corpus, i.e. the marked ancient Chinese sample (from *ZuoZhuan*), is as follows:

二十一年/t ， /w 春/n ， /w 天王/n 將/d 鑄/v 無射/n

This experiment uses two test sets, test set A and test set B. The corpus in test set A is also from *ZuoZhuan*, but it does not intersect with the corpus in training set. Test set B is a collection of corpora from different ancient books. The word statistics and part-of-speech distribution of the data set are shown in Table 1.

3.2 Datasets without labels

In order to further study the effect of the model in the field of ancient Chinese analysis, we selected some corpora from ancient Chinese and ancient traditional medical books that are quite different from Zuozhuan in sentence pattern and content, such as *ZhaoMingWenXuan* « 昭明文选 » (anthology of literature), *ShangHanLun* « 伤寒论 » (treatise on febrile diseases caused by cold) and *ShuoWenJieZiZhu* « 说文解字注 » (Collected commentaries on the *ShuoWenJieZi*).The corpus selected from *ZhaoMingWenXuan* is mainly fragments of Ci and Fu, such as *LuoShenFu*, *ShangLinFu* and so on. The corpus selected from *ShangHanLun* is mainly the disease conclusion and prescription of ancient Chinese medicine. The corpus selected from *ShuoWenJieZiZhu* is mainly explanatory articles. The specific format and contents are shown in Table 2.

Book	Corpus
ZhaoMingWenXuan	髣髴兮若輕雲之蔽月，飄颻兮若流風之回雪。遠而望之，皎若太陽升朝霞；迫而察之，灼若芙蕖出淥波。
	As obscure as a light cloud covering the moon, as drift as a gust of wind blowing up the snow. From afar, it's shining like the soleil and the glory of the dawn,and upon closer inspection, it looks like hibiscus in the green water.
ShangHanLun	太阳病，得之八九日，如疟状，发热恶寒，热多寒少，其人不呕，清便欲自可，一日二三度发。脉微缓者，为欲愈也。
	Disease of Taiyang, got it for eight or nine days, like malaria, have fever and dread cold, fever is more serious than dreading cold, that one won't vomit and still able to defecate normally, symptoms two or three times a day. If the pulse becomes slightly softer, it's about to heal.
ShuoWenJieZiZhu	除，開也。从阜。取以漸而高之意。余聲，直魚切，五部。
	Chu(除) means open. Fu(阜) as the radicals. It to the effect that higher and higher. Yu(余) as the phonetic indictors. *ZhiYu Qie*. It’s located at the fifth part of the Rhyme categories of Old Chinese.

Table 2: Format and content of some corpus

4. Experimental Design

4.1 Integrated Label Design

The commonly used annotation method for word segmentation is {B, I, E, S}, where B represents the first character of word, E represents the last character word, I represents the middle characters of word when the word length is greater than 3, and S represents the word formation of a single character, for example:

*二**B** 十**I** 一**I** 年**E** 春**S***

The actual labels used in this experiment are obtained by the combination of the tagging method mentioned and part-of-speech. The label examples of training corpus are shown in Table 3.

Character	Label
天	*B-n*
王	*E-n*
將	*S-d*
鑄	*S-n*
無	*B-n*
射	*E-n*
，	*S-w*

Table 3: Training corpus label examples

4.2 Network Model Parameters

In this experiment, four network models were used to, which are *BiLSTM*, *BiLSTM_CRF*, *SIKU-RoBERTa* and *SIKU-RoBERTa_CRF*. All models are tested in the same hardware and software environment. The experimental tool and environment selected for this experiment is *pytorch-1.10.0*, *python-3.8* and *cuda-11.3*. The hardware configuration is GPU: *12G RTX3060*, CPU: *20G 7-core Intel(R) Xeon(R) CPU E5-2680 V4 @ 2.40GHz.*

Super Parameter	Value
embedding_size	128
hidden_size	256
num_layers	2
train_batch_size	32
eval_batch_size	8
learning_rate	0.005
num_train_epochs	20
drop_out	0.5

Table 4: Main super parameters of BiLSTM

The super parameters of both BiLSTM network models are shown in Table 4. And The network models super parameters of RoBERTa is shown in Table 5.

Super Parameter	Value
num_attention_heads	12
hidden_size	768
train_batch_size	64
val_batch_size	8
learning_rate	2.0E-5
num_train_epochs	10
drop_out	0.1

Table 5: Main super parameters of RoBERTa

5. Results Analysis

5.1 Evaluation Indexes

In this experiment, due to the small amount of training data, the training data is randomly divided into training set and Validation set according to 9:1, and the 10 fold cross verification method is used to increase the amount of data, enhance the accuracy of the experiment and reduce the error. The confusion matrix between the predicted value and the real value is shown in Table 6.

Confusion Matrix		Actuality	
		Positive	Negative
Prediction	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Table 6: Confusion Matrix

The commonly used evaluation indexes of deep learning model include *P*(Precision), *R*(Recall) and *F1-score*(harmonic mean). *P* reflects the accuracy of the model prediction, *R* reflects the comprehensiveness of the model prediction, and *F1-score* combines the advantages of the two, which can more objectively evaluate the prediction results of the model. The calculation method of the three evaluation indexes is as follows:

$$P = \frac{TP}{TP + FP} \times 100\%$$

$$R = \frac{TP}{TP + FN} \times 100\%$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\%$$

5.2 Cross Verification

In order to more accurately evaluate the performance of *SIKU-Roberta* model, we use the 10 fold cross verification method to evaluate the model. The results of the *Precision*, *Recall* and *F1-score* of each group are shown in Table 7.

Group	Precision	Recall	F1-score	support
1	87.26	87.50	86.99	12260
2	90.32	89.22	89.30	15319
3	91.92	92.66	92.20	14978
4	91.69	92.24	91.68	14003
5	89.89	90.42	89.95	15405
6	90.42	90.60	90.27	16203
7	90.61	93.01	91.72	14809
8	90.16	92.47	91.21	18198
9	89.98	91.91	90.81	17496
10	91.78	93.04	92.30	16793
Mean	90.403	91.307	90.643	15546

Table 7: The result of cross verification with *SIKU-RoBERTa*

Through the comparative analysis of 10 groups of evaluation indexes of models using different pre-training models, it can be seen that the overall *Precision* of part-of-speech tags using *SIKU-RoBERTa* achieves 90.40%, *Recall* achieves 91.31%, and *F1-score* achieves 90.64%.

5.3 Prediction Results

This experiment uses the three network models mentioned in Chapter 4 to test the sequence label prediction task on the *Test Set A* and *B* mentioned in Chapter 3. The results of the final word segmentation and part-of-speech tagging experiment are shown in Table 8 and Table 9.

Test Set	Model	P	R	F1
Test A	*BiLSTM*	92.31	92.88	92.60
	BiLSTM_CRF	92.99	93.42	93.20
	SIKU-RoBERTa	93.09	94.66	93.87
	SIKU-RoBERTa_CRF	95.47	93.48	94.46
Test B	*BiLSTM*	88.38	86.59	87.48
	BiLSTM_CRF	87.98	84.82	86.37
	SIKU-RoBERTa	86.42	93.64	89.89
	SIKU-RoBERTa_CRF	94.39	86.78	90.43

Table 8: Word segmentation experiment results

5.4 Exploratory Experiment Results

The exploratory experiment used the unlabeled test set mentioned in Chapter 3 to evaluate the *SIKU-RoBERTa* model. Since the test set has no label, we can't show our evaluation results digitally. However, with reference to the opinions of relevant professionals, the experimental results of word segmentation and part-of-speech tagging in the above corpus are not as good as those in *Test Set A* or *Test Set B*. Through the analysis and comparison of the corpus, we believe that there are the following reasons:

- Differences in sentence structure: for example, there are great differences in sentence structure between *ZuoZhuan* and *ShuoWenJieZiZhu*.
- Existence of professional terms: there are a large number of disease names in ancient medical texts, such as "太阳病". And the ancients used the inverted phonetic notation, such as "直魚切".
- Difficulty in tagging function words: function words in ancient Chinese are different from those in modern Chinese and English in function and meaning.
- Particularity of poetry: ancient poetry and ancient prose are also different in grammar and semantics.

Test Set	Model	P	R	F
Test A	*BiLSTM*	85.71	86.23	85.97
	BiLSTM_CRF	87.03	87.75	87.39
	SIKU-RoBERTa	88.24	89.73	88.97
	SIKU-RoBERTa_CRF	91.02	89.12	90.06
Test B	*BiLSTM*	73.60	72.11	72.84
	BiLSTM_CRF	75.87	73.14	74.48
	SIKU-RoBERTa	80.33	87.04	83.55
	SIKU-RoBERTa_CRF	88.17	81.06	84.46

Table 9: POS tagging experiment results

6. Conclusion and Discussion

The comparative experiments of four natural language models *BiLSTM*, *BiLSTM_CRF*, *SIKU-RoBERTa* and *SIKU-RoBERTa_CRF* verify that the pre-training model *SIKU-RoBERTa* can improve the accuracy of word segmentation and part-of-speech tagging in ancient Chinese, perform more prominently in the non-specific corpus, and have better generalization ability.

Inspired by the exploratory experiment, there are two thoughts on how to improve the prediction accuracy of the model :

- Expand the training set: increase the diversity of sentence patterns in the training set corpus, so that the model can learn more sentence structures.
- Increase the number of labels: identify some proper nouns through labels.

7. Acknowledgements

Inner Mongolia Autonomous Region university network security and Education Management Information Engineering Research Center construction support project.

8. Bibliographical References

Chang, L., Dongbo, W., Tian, H. H., Qin, Z. Y., & Bin, L.(2021). Research on automatic word Segmentation of Classic Books with external features for digital humanities: A case study of sikuBERT pre-training model. *Library Tribune*, 1-13.

Dongbo, W., & Chang, L. (2021). SikuBERT and SikuRoBERTa: Research on the construction and application of pre-training model of *SiKuQuanShu* for Digital Humanities. *Library Tribune*, 1-14.

RunHua, X., & Xiaohe, C. (2021). A Method of Segmentation on "Zuo Zhuan" by Using Commentaries. *Journal of Chinese Information Processing, 26*(02), 13-17+45.

Yundong, G., Yiqin, Z., Huan, L., & Dongbo, W. (2021).

Automatic part-of-speech tagging of Chinese ancient classics in the context of digital humanities research:A case study of SIKU-BERT pre-training model. *Library Tribune* 1-11.

Zhiting, Y., & Hanjie, M. (2021). Automatic—annotation me thod for e me rge ncy te xt corpus based on BE R T. *Intelligent Computer and Applications.*

Ning, C., Bin, L., Sijia, G., Xingyue, H., & Minxuan, F. (2020). A Joint Model of Automatic Sentence Segmentation and Lexical Analysis for Ancient Chinese Based on BiLSTM-CRF Model. *Journal of Chinese Information Processing, 34*(04), 1-9.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, jun). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, Minneapolis, Minnesota.

Dongbo, W., Shuiqing, H., & lin, H. (2017). Researches of Automatic Part-of-speech Tagging for Pre-Qin Literature Based on Multi-feature Knowledge. *Library and Information Service, 61*(12), 64-70. doi:10.13266/j.issn.0252-3116.2017.12.008

TONG FEI C，WEI MENG Z，XUE QIANG L，et al.A kalman filter based human-computer interactive word segmentation system for ancient chinese texts[M]. Chinese computational linguistics and natural language processing based on naturally annotated big data.Berlin，Heidelberg：Springer，2013：25-35.

FANG M，JIANG Y，ZHAO Q，et al.Automatic word segmentation for Chinese classics of tea based on tree-pruning[C]//2009 Second International Symposium on Knowledge Acquisition and Modeling. IEEE, 2009，（01）：438-441.

Jiannian, H. (2009). *Research on Automation of Sentence Segmentation, Punctuation and Word Segmentation of Agricultural Ancient Books.* (D). Nanjing Agricultural University, Available from CNKI.

9. Language Resource References

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