Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 155–158
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC 4.0

Construction of Segmentation and Part of Speech Annotation Model in Ancient Chinese

Longjie Jiang, Qinyu Chang, Huyin Xie, Zhuying Xia
NANJING NORMAL UNIVERSITY ZHONGBEI COLLEGE
Zhenjiang, Jiangsu, China
wppwlp010820@163.com, {1225048113, 1963912428, 2900997927}@qq.com

Abstract

Of the four ancient civilizations, China is the only one whose history has never been interrupted over the past 5000 years. An important factor is that the Chinese nation has the fine tradition of sorting out classics, recording history with words, inheriting culture through continuous collation of indigenous accounts, and maintaining the spread of Chinese civilization. In this research, the siku-roberta model is introduced into the part-of-speech tagging task of ancient Chinese by using the data set of Zuozhuan, and good prediction results are obtained.

Keywords: Natural Language Processing, Old Chinese, Word Segmentation, POS Tagging

1. Introduction

Chinese classics are vast and profound. From ancient oracle bone inscriptions to books written on paper, they have a long history of more than 3,000 years. They are numerous, diverse in form and rich in content. These classics are important civilization achievements created by the Chinese nation in the long history, and reflect the Chinese people's thought, literature, art, science and technology.

However, due to the grammatical characteristics of ancient Chinese, the use of words and other words differs greatly from modern Chinese. Digging out the essence of information from the ancient Chinese treasure house has become a huge problem. In recent years, researches on word segmentation and part-of-speech tagging of modern Chinese have achieved fruitful results, while those on ancient Chinese are still insufficient.

The usage of words in ancient Chinese is flexible, with many concurrent words and flexible parts of speech, i.e. most sequences have different segme.

2. Correlation Study

2.1 Study on Part of Speech Labeling in Ancient Chinese

Such system of ancient Chinese has experienced thousands of years of development. A word has unique significance in different times and contexts. According to different historical periods, ancient Chinese can be divided into ancient Chinese, medieval Chinese and modern Chinese. Therefore, it is not feasible to train the ancient Chinese model that is similar to the prediction of modern Chinese. Due to the different standards of part-of-speech tagging, it is also infeasible to train the ancient Chinese word segmentation model and the ancient Chinese part-of-speech tagging model with ancient Chinese corpus in different periods, which will cause trouble in the process of training supervised learning model based on corpus.

Part-of-Speech Tagging refers to assigning unique part-of-speech tags to each word's segmentation in the text according to certain marking rules, such as adjectives, nouns, verbs, etc. The labeling method is as follows:

Table 1: Gender of word m2arkers

Number	Tagging	POS	Number	Tagging	POS
1	n	普通名词	11	p	介词
2	nr	人名	12	c	连词
3	ns	地名	13	u	助词
4	t	时间名词	14	d	副词
5	v	动词	15	y	语气词
6	gv	古代动词	16	s	拟声词
7	a	形容词	17	j	兼词
8	m	数词	18	w	标点
9	q	量词	19	i	词缀
10	r	代词			

Specific annotation samples are as follows:

未/d 王命/n ，/w 故/c 不/d 書/v 爵/n ，/w 曰/v ：/w “/w 儀父/nr ”/w ，/w 貴/sv 之/r 也/y 。/w

In this study, the punctuation of *Zuozhuan*, word segmentation and word class label text are used as training data packets. *Zuozhuan* is an ancient Chinese masterpiece in the Spring and Autumn Period (770－476 BC), which is believed to be dated back to the Warring States Period (475–221 BC). *Zuozhuan* is a comment on the history of Spring and Autumn Period (770–476 BC).

The training data were distributed according to Nanjing Normal University's Ancient Chinese Word Segmentation and Corpus Guide. According to this format, annotations are encoded in UTF-8 plain text files. There is no word boundary in Chinese text. Therefore, the original text contains characters and punctuation marks. After manual annotation, text boundaries and part-of-speech tags are added to the text. As shown in Table 1, each word has a POS tag in the form of Word / POS. Each word is separated by a space, and punctuations are also treated as words.

Test data is provided in original format and only Chinese characters and punctuation are provided. There are two test data sets. Test A is designed to see how the system runs the data in the same name book. Zuozhuan_Test is extracted from *Zuozhuan* and has no overlap with the Zuozhuan_Train. Zuozhuan_Test does not allow the team to use it as training data. Test B aims to explore how the system processes similar data (texts with similar contents but from different books), the size of which is similar to Zuozhuan Test.

2.2 Sequence Annotation Studies Based on Deep Learning

Deep learning is a kind of machine learning, which simulates the mechanism of human brain to explain and analyze the data of image, speech and text by establishing deep neural network. Different from the traditional statistical-based machine learning model, deep learning attempts to automatically complete feature extraction. In recent years, it has received extensive attention in the field of natural language processing and has achieved remarkable progresses in application research. Since part-of-speech tagging of word segmentation can be regarded as one of the sequence tagging tasks, the following reviews the related research based on the models involved in sequence tagging.

3. Construction of Model

3.1 Model Introduction

(1) FLAT + Sikuroberta

1. Using Lattice framework. FLAT proposed by Fudan University are adopted as the subject of lexical enhancement.

2. Switching of pre-training model. The bert-wwm originally used by FLAT is replaced with the Sikuroberta 2.0 pre-training model of closed test.

3. Training of word vectors. 50-dimensional unigram, bigram and word-level word vectors are trained based on word segmentation data from the ' Sikuquanshu ' History Department.

FLAT + Sikuroberta model is constructed based on the above three steps.

(2) FLAT

In ACL 2020, the research team of Xipeng Qiu in Fudan University proposed FLAT: Chinese NER Using Flat-Lattice Transformer. FLAT has two innovations. First, it designs a position encoding based on Transformer to fuse Lattice structure, which can introduce lexical information losslessly. Second, it integrates the dynamic structure of lexical information based on Transformer, supports parallel computing, and greatly improves inference speed. FLAT reconstructs the original Lattice, and cleverly designs position encoding to fuse Lattice structure. Each character and vocabulary is constructed two head position encoding and tail position encoding, so that FLAT can directly model the interaction between characters and all matching vocabulary information. FLAT uses relative position coding to make Transformer suitable for NER tasks.

$$\begin{aligned} A_{i,j}^{*} = {} & W_q^T E_{x_i}^T E_{x_i} W_{K,E} \\ & + W_q^T E_{x_i}^T R_{i,j} W_{K,R} \\ & + u^T E_{x_i} W_{k,E} \\ & + u^T R_{ij} W_{k,R} \end{aligned}$$

Four relative distances are proposed to represent the relationship between xi and xj, including the relationship between characters and words.

$$d_{ij}^{(hh)} = head[i] - head[j]$$

$$d_{ij}^{(ht)} = head[i] - tail[j]$$

$$d_{ij}^{(th)} = tail[i] - head[j]$$

$$d_{ij}^{(tt)} = tail[i] - tail[j]$$

$d_{ij}^{(hh)}$ represents the head distance from head of xi to xj, which is similar to that of xi.The relative position encoding is expressed as :

$$R_{ij} = ReLU\left(W_r\left(P_{d_{ij}^{(hh)}} \oplus P_{d_{ij}^{(ht)}} \oplus P_{d_{ij}^{(th)}} \oplus P_{d_{ij}^{(tt)}}\right)\right)$$

$d_{ij}^{(hh)}$ calculation method is the same as vanilla Transformer :

$$P_d^{(2k)} = \sin\left(d/10000^{2k/d_{model}}\right)$$

FLAT vocabulary enhancement uses transformer to design a positionencoding to fuse the Lattice structure, efficiently introduce vocabulary information, and fuse the dynamic structure of vocabulary information, which can capture long-distance dependence and greatly improve inference efficiency.

(3) Sikuroberta

The figure below illustrates the training process of the Sikuroberta model

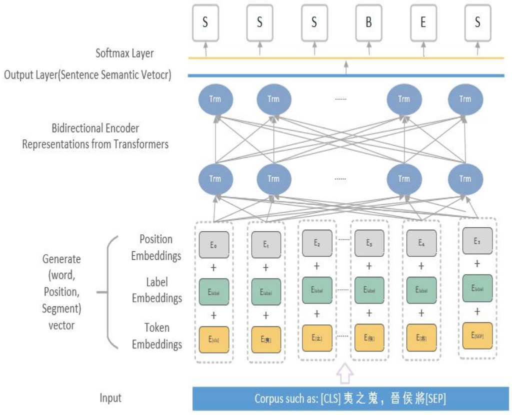
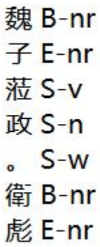


Figure 1: Training process of Sikuroberta

As shown in the above figure, at the Embedding layer, the BERT model divides the input Chinese sequences in words and maps the characters into numerical sequences using its own Chinese dictionary. For example, when the model reads into the sequence of "夷之蒐， 晉侯將", this sentence is first divided by BERT model into characters with sequence start mark [CLS] and termination mark [SEP], and converted into input sequence[CLS]，夷，之，蒐，，，晉，侯，將，[SEP].Then it automatically combines the corresponding index value of each word to generate the word vector, the position of the word in the sentence, and the segment vector representing the sentence category, together generate a combined vector. Through the stacked multi-layer bidirectional transformer encoder, the final result through a softmax layer can obtain the maximum probability of each character, and the sequence annotation can be realized by exporting this series of labels. It is fairly suitable for discriminative tasks such as text classification and sequence annotation, and it is one of the most popular models in the NLP industry. In this experiment, we selected the Sikuroberta model provided by EvaHan2022. This model has completed the pre-training on the punctuation-free ‘ Four Library Encyclopedia ’ which removes the annotation information, and has remarkable effect on the Chinese natural language processing task.

3.2 Corpus Processing and Experiment

Combined with lexical information, the tag carries the dual information of word segmentation and part of speech. The experiment uses five-word tagging set, namely { B, M, E, S, O }, B represents the first word in the beginning, M represents the middle word, E represents the end word, S represents the word, O represents non-entity. After combining part-of-speech information, the labeled samples are shown in Figure 2.



魏	B-nr
子	E-nr
蒞	S-v
政	S-n
。	S-w
衛	B-nr
彪	E-nr

Figure 2: Examples of corpus preprocessing

In the above graph, it can be seen that ' 魏 ' is the first word in the beginning of this word segmentation, and the part of speech is nr; ' 子 ' is the end word of this word segmentation, and the part of speech is nr. ' 蒞 ', ' 政 'and ' . 'are single words.

3.3 Experimental Environment

The model is built based on Pytorch and FastNLP framework. The NVIDIA card is configured as follows :

Table 2:NVID card configuration

CUDA Version	GPU	Memory
10.2	NVIDIA Tesla P100	32GB

The table above shows the equipment used in this model construction.

3.4 Parameter Regulation

The following is the hyperparameter setting for the various models to reach the optimal state.

Table 3: The hyperparameters of the model

Module	Parameters

word2vec	sg=1,size=50,min_count=1,
	workers=15,sample=1e-3

Sikuroberta	epoch=30,batch-64,
	learning-rate=2e-5

Sikuroberta +CRF	epoch=30,batch=64,
	learning-rate=15e-5

FLAT +Sikuroberta	bert_lr_rate=0.0.5,
	embed_lr_rate,batch=25,
	epoch=50,fix_bert_epoch=20,
	max_seq_length=61

3.5 Model effect comparison and analysis

The study used three models for pos tagging of data, and compared their accuracy, recall rate and F score to select the best model. The three models are Sikuroberta, Sikuroberta + CRF and FLAT + Sikuroberta. We divided the training data by 9:1 and showed the results of pos tagging obtained by the three models.

Table 4: POS tagging results of the three models

Sikuroberta	Precision	Recall	F score
	90.21	90.56	90.37
Sikuroberta	Precision	Recall	F score
+CRF	85.97	85.77	85.87
FLAT	Precision	Recall	F score
+Sikuroberta	91.32	91.20	91.26

The figure shows the pos tagging results of the three models. The F score of Sikuroberta model can reach 90.37 %, and the F score of Sikuroberta + CRF model is only 85.87 %. The performance of Sikuroberta is higher than that of Sikuroberta + CRF. After adding CRF layer to BERT, the F score is not improved. In order to further improve the performance of the model, the present research used FLAT + BERT proposed by Qiu Xipeng team of Fudan University. In order to make FLAT adapt to ancient Chinese, the study replaced the bert-wwn model used in modern Chinese in FLAT with Sikuroberta in closed test. Using word2vec to train the 50-dimensional word vector.The F score of FLAT + Sikuroberta model can reach 91.26 %, which is 0.89 % higher than that of Sikuroberta, and the recall rate is 0.64 % higher, which further improves the overall performance of the model.

3.6 Data test results

Through the research, we selected the FLAT+ Sikuroberta model as the final model to obtain the prediction data of it. Based on the prediction data testa and testb released by EvaHan2022, we used FLAT+Sikuroberta model to predict data and got a best result. As for the test data testa and testb released by EvaHan2022, testa and training data are from the same book, while testb and training data are not in the same book but similar in content.We also used the FINAL script as a scorer to obtain scores. The final test data format and results are shown as follows:

孟懿子/nr 會/v 城/v 成周/ns ，/w 庚寅/t ，/w 栽/v 。/w

Table 5: Results of pos tagging

	Precision	Recall	F score
Testa_closed	88.79	87.54	88.16
Testa_open	88.97	86.48	87.70
Testb_open	89.69	89.25	89.47

Table 6: Results of word segmentation

	Precision	Recall	F score
Testa_closed	92.75	91.44	92.09
Testa_open	92.77	90.17	91.45
Testb_open	95.26	94.78	95.02

From Table 5 and Table 6, in testa of the closed mode, the score of word segmention F1 is 92.09%, and pos tagging F1 is 88.16%. In testa of open mode, the the score of word segmention F1 is 91.45%, and pos tagging F1 is 87.70%. In testb of open mode, the score of word segmentation F1 is 95.02%, and pos tagging F1 is 89.47%.

4. Conclusion

Under the development prospect of artificial intelligence and digital humanities, the research on ancient Chinese is relatively weak. Therefore, the result of pos tagging of ancient books is of great help to the subsequent research, such as the study of Ancient Chinese Literature Search, historiography, philology and Chinese history. Based on the learning model FLAT + Sikuroberta, this paper constructed the pos tagging pattern of ancient Chinese. In tasta under the same book, the score of pos tagging F can reach 87.70%, the score of word segmentation F can reach 91.45%. In testb with similar contents under different books, the score of pos tagging F can reach 89.47%, the score of word segmentation F can reach 95.02%. It can be successfully applied to pos tagging and word segmentation, has achieved the practical goal.

5. References

[1]Liu Chang, Wang Dongbo, Hu Haotian, Zhang Yiqin, Li Bin. Dictionary of integrating external characteristics for digital humanities .Research on automatic word segmentation - Taking sikuBERT pre-training model as an example [J / OL].Library forum.

[2]Zhang Qi, Jiang Chuan, Ji Youshu,et al. Unified Model for Word Segmentation and POS Tagging of Multi-Domain Pre-Qin Literature[J]. Data Analysis and Knowledge Discovery, 2021, 5(3): 2-11.

[3]Xiaonan Li, Hang Yan, Xipeng Qiu , Xuanjing Huang.FLAT: Chinese NER Using Flat-Lattice TransformerFLAT: Chinese NER Using Flat-Lattice Transformer[C].ACL2020, 2020.

[4]Hu Jie, Hu Yan, Liu Mengchi, Zhang Yan.Chinese named entity recognition based on knowledge base and entity enhanced BERT model[J/OL].Journal of Computer Applications.2021.

[5]Yue Zhang,Jie Yang.Chinese NER Using LatticeLSTM[J].arXiv:1805.02023 [cs.CL].2018.