Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 52–58
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Integration of Automatic Sentence Segmentation and Lexical Analysis of Ancient Chinese based on BiLSTM-CRF Model

CHENG Ning[1], LI Bin[1,2], XIAO Liming[1], XU Changwei[1], GE Sijia[1], HAO Xingyue[1], FENG Minxuan[1]
1.School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China
2. Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA
chengninmo@foxmail.com, libin.njnu@gmail.com, lmxiao1leo@gmail.com, changweixu36@gmail.com,
sijiage007@gmail.com, haoxingyue@hotmail.com, fengminxuan@njnu.edu.cn

Abstract

The basic tasks of ancient Chinese information processing include automatic sentence segmentation, word segmentation, part-of-speech tagging and named entity recognition. Tasks such as lexical analysis need to be based on sentence segmentation because of the reason that a plenty of ancient books are not punctuated. However, step-by-step processing is prone to cause multi-level diffusion of errors. This paper designs and implements an integrated annotation system of sentence segmentation and lexical analysis. The BiLSTM-CRF neural network model is used to verify the generalization ability and the effect of sentence segmentation and lexical analysis on different label levels on four cross-age test sets. Research shows that the integration method adopted in ancient Chinese improves the F1-score of sentence segmentation, word segmentation and part of speech tagging. Based on the experimental results of each test set, the F1-score of sentence segmentation reached 78.95, with an average increase of 3.5%; the F1-score of word segmentation reached 85.73%, with an average increase of 0.18%; and the F1-score of part-of-speech tagging reached 72.65, with an average increase of 0.35%.

Keywords: sentence segmentation of ancient Chinese, word segmentation, part-of-speech tagging, BiLSTM-CRF, ancient Chinese information processing

1. Introduction

Lexical analysis is the most basic task of Chinese information processing, including automatic word segmentation, part of speech tagging, and named entity recognition. Besides the above tasks, the basic task of information processing in ancient Chinese also includes automatic sentence segmentation. Chinese ancient books have a vast number of texts, and most of them are unpunctuated, which brings great difficulties for readers to read and study. The use of advanced natural language processing technology for automatic sentence segmentation and lexical analysis of ancient Chinese can not only facilitate readers to read, but also of great significance to the arrangement of ancient books, the development of ancient Chinese and the intelligent application of ancient Chinese.

Most of the research on information processing in ancient Chinese is focused on a specific subtask, such as automatic sentence segmentation and word segmentation, part of speech tagging and named entity recognition. To complete the basic task of ancient Chinese information processing, most scholars adopt different research methods and techniques, and each subtask need to be completed in turn, which greatly affects the processing efficiency of the machine. Moreover, using sentence segmented by machine to go on doing word segmentation and part of speech tagging are easy to result in multi-level diffusion of tagging errors, which affects the accuracy of overall tagging task.

In this paper, a tagging system integrating automatic sentence segmentation and lexical analysis in ancient Chinese is designed and completed. BiLSTM-CRF model is used to joint learn sentence segmentation, word segmentation and part of speech information. Due to the relative shortage of tagged ancient Chinese corpus, most of the previous studies were conducted according to a special book, and the corpus scales were relatively small, so the training model could not be well applied to other types of ancient Chinese texts. Based on the existing resources, this paper constructs four kinds of annotated corpus written in different ages, and verifies the effect of the integrated annotation on different test sets by using the neural network model.

2. Model introduction

RNN model and its variants, which are suitable for sequence tagging, have greatly changed the research methods of natural language processing. RNN can be regarded as a multiple overlay structure of the same network. It performs the same operation for each element in the sequence, and each operation depends on the previous calculation results. In theory, RNN can use any length of sequence information, but in practice, only some previous steps can be reviewed. LSTM neural network is a kind of special RNN. Based on the original RNN model, input gate, forgetting gate and output gate are added. Neurons will selectively forget the useless information for current output. It inherits the advantage that RNN can keep the preorder's information, and overcomes the problem that RNN can't really capture the long-distance dependency in the text.

BiLSTM is a model put forward by Schuster in 1997 to solve the problem that LSTM can't retain the post information. The main idea of the model is to set up two LSTM structures in the front and back direction of the training sequence. By splicing the LSTM in two directions to capture the preorder and post order's information, the information in the whole training sequence can be retained to the greatest extent.

The BiLSTM-CRF model structure used in this paper was first proposed by Huang et al. The output of BiLSTM layer is a probability matrix, which is calculated by BiLSTM based on the optimal result of each moment. In this way, the output tag doesn't consider the influence of the previous tag. For example, the word "孟子" appears in the input sequence "孟子*(name)*卒*(die)*继室*(second wife)*以*(a conjunction)*馨子*(name)*", in which "孟" is the first character and "子" is the last character. The model may predict both "孟" and "子" as the first character, such situation should be avoided in the lexical analysis task of ancient Chinese. CRF is a framework for an undirected graph model that can be used to define the joint probability distribution of a tag sequence in a situation that a set of observed sequences need to be tagged. Assume that X is the random variable of the data sequence to be annotated, and Y is the random variable of the corresponding tag sequence. For example, X is the set of sentences in natural language, and Y is the part of speech set that used to mark these sentences. Random variables X and Y are jointly distributed and a conditional model P(Y|X) is constructed according to the pairs of observation sequence and label sequence. The CRF layer is matched with the output layer of BiLSTM, so that the output sequence of BiLSTM becomes the observation sequence of CRF, and then CRF calculates the optimal solution of the whole sequence in probability without ignoring the interaction between sequence element tags.

3. Construction of corpus

Ancient texts were selected according to different historical stages, and the corpus with the same size was extracted from the traditional version of *Tso Chuan* (左傳, Han dynasty, 722BC~468BC), *Brush Talks from Dream Brook* (夢溪筆談, Song dynasty, AD1086~AD1093), *Fantastic Tales by Ji Xiaolan* (閱微草堂筆記, Qing dynasty, language style is more colloquial, AD1789~AD1798), and *Documents of History of Qing Dynasty* (清史稿, Republic of China, AD1914~AD1927) as the experimental data set of this paper. The purpose of constructing a corpus by age is to explore the generalization ability of the model for text annotation in different ages after training based on mixed corpus of different ages. The data set is manually proofread on the basis of machine-assisted word segmentation and POS tagging. Kappa was used for labeling consistency test and the Kappa value was higher than 0.8, indicating a higher degree of labeling consistency. The specification of POS tags refers to Ancient Chinese Corpus published by LDC[1], totaling 21 tags. The experimental data set is divided into training set, development set and test set according to the ratio of 8:1:1. Among them, the training set is a mixed corpus composed of 80% of the corpus in *Tso Chuan*, *Brush Talks from Dream Brook*, *Fantastic Tales by Ji Xiaolan*, and *Documents of History of Qing Dynasty*. Based on this mixed corpus, this paper discusses the annotation ability of the model to texts of various ages. The experimental corpus set “：，。；！？” six kinds of punctuation as sentence breaks, and each text sequence divided by two sentence breaks is treated as a sentence, with all other punctuation ignored. Table 1 is a general overview of the experimental data set.

4. Integrated word position tag design

Xue is the first to put forward a character-based learning method of sequential annotation, who uses four kinds of tags, which is LL(stands for left boundary of a word), LR(stands for monosyllabic word), MM(stands for the middle of a word) and RR(stands for the right boundary of a word), to express the segmentation and annotation information of characters, thus it translates word segmentation task into serialized annotation task formally for the first time.

The data set	The training set			The development set			The test set		
	#character	#word	#sentence	#character	#word	#sentence	#character	#word	#sentence
Tso Chuan	75,000	65,000	15,000	9136	7755	1917	9280	7738	2046
Brush Talks from Dream Brook	81,000	63,000	13,000	9483	8384	1662	9825	8378	1643
Fantastic Tales by Ji Xiaolan	81,000	69,000	14,000	9722	8699	1745	9789	8680	1784
Documents of History of Qing Dynasty	81,000	57,000	12,000	10248	8851	1651	9991	8159	1432
Total	32,400	25,400	54,000	38,000	34,000	6975	38,000	33,000	6905

Table 1 : Experimental data set

This paper uses this method of character annotation to construct an ancient Chinese integrated-analysis annotation system. For this model, the problem is actually a tag multi-classification problem, where each character needs to be assigned to a specific tag type.

Word segmentation layer (WS): Using B, I, E, S four tags. B means that the current character is at the beginning of a multi-character word. I means that the current character is at the middle of a multi-character word. E means that the current character is at the ending of a multi-character word. S represents the current character is a one-character word. After transforming the character annotation sequence, the sentence segmentation results can be calculated out. For example:

Character annotation: 九 B 月 E ， S 晉 B 惠 I 公 E 卒 S 。 S 懷 B 公 E 立 S ， S
After the transformation: 九月*(September)* ， 晉惠公 卒*(die)* 。懷公 立*(ascend the throne)* ，

POS tagging layer (POS): Tagging the part of speech of the word to which each character belongs. Meanwhile, incorporating physical tags (personal name *nr*, place name *ns*) into POS. Then, adding POS on the basis of WS so that each character can corresponds to its position in the word

[1] LDC Ancient Chinese Corpus https://catalog.ldc.upenn.edu/LDC2017T14

and the part of speech it represents or entity information it has.

九 B-t 月 E-t，S-w 晉 B-nr 惠 I-nr 公 E-nr 卒 S-v 。S-w 懷 B-nr 公 E-nr 立 S-v，S-w

Each character is tagged word segmentation tag and POS tag, connected by “-“.Take “晉 B-nr 惠 I-nr 公 E-nr” as example, “晉” is the first character of a personal name, “惠” is a character in the middle of a personal name, “公” is the last character in a personal name, so that “晉惠公” can be segmented and recognized to a person's name, whose reality tag is represented as “nr”.

Sentence segmentation layer (SS): Tagging whether a character is at the end of a sentence. Adding SS layer on the basis of WS and POS, so that each character can be corresponded with three layers, i.e., word segmentation, part of speech and sentence segmentation.

九 B-t-O 月 E-t-L 晉 B-nr-O 惠 I-nr-O 公 E-nr-O 卒 S-v-L 懷 B-nr-O 公 E-nr-O 立 S-v-L

If a character in the corpus is at the break of a sentence, such as “月”, “卒” and “立” in the sentence, then tag “L” will be put after the part of speech tag, otherwise, tag “O” will be put after the part of speech tag.

During the process of corpus preprocessing, three-layers tags categories (WS, POS, SS) can be processed in different ways:

WS+POS+SS (e.g., 卒 S-v-L) is a three-layers tag. Under this annotation level, the annotation effect of each subtask, such as sentence segmentation (SS), can be calculated.

There is WS+POS (e.g., 卒 S-v) in two-layers tags. Under this annotation level, the effects of word segmentation (WS) and POS tagging (WS+POS) can be calculated.

There is WS (e.g., 卒 S) and SS (e.g., 卒 L) in one-layer tags. The effect of sentence segmentation or word segmentation can be calculated.

5. Evaluation indexes

The experimental training set is used for feature learning and training of the model, and the test set is used to verify the results of automatic tagging. For the evaluation of automatic tagging results, F1-score (harmonic mean), the most commonly used evaluation index in sequence tagging, is used to measure the effect of the model. F1-score is calculated from P(precision) and R(recall), and the calculation formula is:

$$\mathrm{F1} = \frac{2 * \mathrm{P} * \mathrm{R}}{\mathrm{P} + \mathrm{R}}$$

The calculation of Precision is as follows:

$$\mathrm{P} = \frac{\text{Correct number of tags}}{\text{Number of machine tags}}$$

The calculation of Recall is as follows:

$$\mathrm{R} = \frac{\text{Correct number of tags}}{\text{Number of all tags in the corpus}}$$

Based on the above evaluation metrics, sentence segmentation, word segmentation, part of speech tagging results are calculated. Sentence segmentation calculation is based on sentence rather than characters, that is, according to the label “L”. If both machine and manual tagging results are “L”, it is correct. Word segmentation and part of speech are calculated on the basis of words rather than characters. Taking POS tagging as example, it is assumed that the word 孟子(Mencius) is predicted as “孟S-nr子S-nr”. Although the model gets a correct part of speech based on characters, however, the word segmentation is wrong, and the correct answer should be “孟B-nr子E-nr”. To determine whether a word belongs to the correct part of speech, whether the character is correctly divided into words should be determined first, that is, determination should be based on the correct word segmentation.

6. Experimental design and result analysis

The results of Experiment 1 are the super parameters obtained by manual parameter adjustment on the development set, and the results of Experiment 2, Experiment 3 and Experiment 4 are obtained on the test set.

Experiment 1 will verify the necessity of adding word vectors into the integration analysis of ancient Chinese and investigate the effect of word vectors of different dimensions on the results of integrated annotation. Generally speaking, the higher the dimension of the word vector, the more semantic features it contains, but they are not absolute positively correlated. Based on nearly 1.5 billion characters of traditional ancient Chinese raw corpus (from Imperial Collection of Four and other ancient Chinese corpus), selecting word2vec as the tool, CBOW (Continuous Bag of-Words Model) as the model, we carry out character vector pretraining. The experiment sets the word vector dimension to 50, 100, 128 and 200 respectively, selects *Tso Chuan* test set as the test corpus, and adopts “WS+ POS+ SS” as its tagging layer, which is a tagging method of integrating sentence segmentation and lexical analysis. By manually adjusting parameters on the development set, the final hyper-parameter adopted is shown in Table 2.

Word vector dimension	50/100/128/200
Number of hidden layers	1
Number of hidden units	200
Minimum number of samples	64
Dropout rate dropout	0.5
The optimizer	Adam
Learning rate	0.001

Table 2: Experimental hyper-parameter setting

In the BiLSTM-CRF structure, based on experiments on the development set, it is found that the number of layers in BiLSTM had little influence on the precision, so the number of hidden layers in the model, namely the number of layers in BiLSTM, is set as 1. The number of hidden nodes in the sequence tagging task is usually from 200 to 600, and 200 is taken as the parameter here. The minimum sample size is set to 64, with each sample size controlled between 50 and 60. The optimization of the model adopts the “Adam” algorithm, which has a good effect in the sequence tagging task. The Dropout method is used to reduce overfitting. A Dropout with a parameter of 0.5 is added between the BiLSTM layer and the full connection layer, which can weaken the excessive interaction between various features caused by the small amount of data, so that the model has the optimal generalization ability and the lowest degree of overfitting. The experimental results are shown in Table 3.

Word vector dimension	Sentence segmentation	Word segmentation	POS tagging
No word vector	82.16	88.23	78.36
50 dimensions	83.07	89.39	79.53
100 dimensions	83.89	90.19	80.59
128 dimensions	**84.11**	**90.24**	**80.88**
200 dimensions	83.58	89.83	80.42

Table 3: The F1-score of integration of sentence segmentation and lexical analysis(unit %)

As can be seen in table 3, the addition of word vector is necessary for sentence segmentation and lexical analysis tasks in ancient Chinese, especially for POS tagging tasks, which increased by 2.5 percentage points. In the word vector dimension setting, the experiment shows that 128 dimensions is the best for the integrated automatic tagging of ancient Chinese. In order to verify the training effect of the word vector under this dimension, cosine similarity is used to calculate the semantic correlation between the two word vectors: Assume word vector A=(A1,A2,…,An), B=(B1, B2,…,Bn), the formula for cosine similarity is as follows:

$$\cos\theta = \frac{\sum_{1}^{n}(\mathrm{Ai} \times \mathrm{Bi})}{\sqrt{\sum_{i=1}^{n}(\mathrm{Ai})^2} \times \sqrt{\sum_{i=1}^{n}(\mathrm{Bi})^2}}$$

i represents the dimension of the vector, and *Ai* represents the specific value of the *i*-dimension of the character *A*. Taking characters 也 *(modal particle)* and 曰 *(say)* as examples, the calculation results are as follows in Table 4:

The most semantically relevant word of 也	The most semantically relevant word of 曰
矣*(modal particle)* 0.662	云*(say)* 0.696
之*(modal particle)* 0.659	謂 0.584
乎*(modal particle)* 0.658	也 0.514
謂*(say)* 0.652	言*(say)* 0.500
非*(be not)* 0.593	問*(ask)* 0.465
歟*(modal particle)* 0.584	耶 0.434
耶*(modal particle)* 0.571	荅*(answer)* 0.415
哉*(modal particle)* 0.563	答*(answer)* 0.413
以*(with)* 0.525	為*(do)* 0.412

Table 4: Semantic relevancy calculation results

In experiment 2, for testing the performance of BiLSTM-CRF model in tagging ancient texts, we used IDCNN (Iterated Dilated Convolutions) and non-CRF-layer BiLSTM model to compare with it. DCNN (Dilated Convolutions) was first proposed by (Yu et al., 2015) and applied to image semantic classification. IDCNN model structure is generated based on DCNN. Drawing on the advantages of CNN and RNN, IDCNN takes into account the parallel processing and breadth of context feature extraction, so it is also widely used in sequence tagging tasks. In this experiment, *Tso Chuan* is chosen as test set, and tagged in the method of integrating sentence segmentation and lexical analysis. Keeping other experimental variables (e.g., training corpus, word vector dimension) consistent, we investigate tagging effect of different models in the word segmentation task under integrated tagging layer. The experimental results are shown in Table 5.

Neural network models	*Tso Chuan* testing set (unit %)		
	P	R	F1
IDCNN	88.25	89.28	88.76
BiLSTM	**89.39**	90.05	89.71
BiLSTM-CRF	89.37	**91.13**	**90.24**

Table 5: Word segmentation performance of different models on *Tso Chuan*

The results of comparative experiments show that in ancient Chinese word segmentation task, the precision of BiLSTM-CRF model is only 0.02% lower than BiLSTM model, which is almost not different, and the recall is 1.08% higher than non-CRF-layer BiLSTM model, and F1-score is 1.48% higher than IDCNN and 0.53% higher than BiLSTM. As a result, BiLSTM-CRF model's performance is generally higher than IDCNN model and BiLSTM model in ancient Chinese word segmentation task.

This experiment was not carried out in the other three books, but the effects should be good because of the BiLSTM-CRF's advantage compared to the other two models.

The third set of experiments focuses on four kinds of texts, including *Tso Chuan*, *Brush Talks from Dream Brook*, *Fantastic Tales by Ji Xiaolan*, and *Documents of History of Qing Dynasty*. In each text's in-domain experiment, the training and testing corpus we used are both from the same text. The purposes of experiment 3 is to explore the modeling ability of the model that integrates sentence segmentation and lexical analysis applying to various texts, and to compare the result with experiment 4 which based on mixed corpus.

Tagging layer in the experiment is "WS+POS+SS", i.e., the tagging method of integrating sentence segmentation and lexical analysis. The experimental parameters are consistent with the previous ones. The experimental results are shown in Table 6.

Tagging layers		***Tso Chuan***			***Brush Talks from Dream Brook***			***Fantastic Tales by Ji Xiaolan***			***Documents of History of Qing Dynasty***		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
integration	sentence segmentation	85.8	83.0	84.4	72.4	67.4	69.8	70.2	71.7	71.0	87.7	87.0	**87.4**
	word segmentation	89.9	92.1	**90.9**	86.8	84.8	85.8	85.8	87.9	86.8	82.8	77.3	80.0
	POS tagging	81.0	83.0	**82.0**	66.7	65.1	65.9	71.1	72.9	72.0	72.7	68.0	70.3

Table 6: Experimental results of BiLSTM-CRF model applying to various texts under “WS+POS+SS” layer (unit %)

Because of the differences in the age and genre of the four texts, the experimental results of the model for each corpus are quite different. By comparing the F1-score of word segmentation task, POS tagging task and sentence segmentation task, we found that in word segmentation task, *Tso Chuan* performances best, *Fantastic Tales by Ji Xiaolan* ranks the second, *Documents of History of Qing Dynasty* is the worst; in POS tagging task, *Tso Chuan* and *Fantastic Tales by Ji Xiaolan* have the same rank as last task, but *Brush Talks from Dream Brook* is the worst; in sentence segmentation task, *Tso Chuan* and *Documents of History of Qing Dynasty*’s effects are relatively good, far more accurate than *Fantastic Tales by Ji Xiaolan* and *Brush Talks from Dream Brook*. After analyzing the model tagging errors, we found that *Brush Talks from Dream Brook* contains a large number of non-repetitive professional terms in various disciplines, for example, in sentence “南呂調皆用七聲*(scales)*：下五、高凡、高工、尺、高一、”, the words “下五”, “高凡” are proper names related to music. The relatively sparse data of proper names makes it difficult for the model to learn the relevant features, which is the main reason that *Brush Talks from Dream Brook* performances worse in POS tagging task.

Experiment 4 is designed from two dimensions: (1) in the horizontal dimension, the experiment discusses the differences of model based on mixed corpus, tagging in different ages’ corpus under a same tagging layer, and investigates the models’ generalization ability considering the result of experiment 3; (2) in the vertical dimension, the experiment compares the tagging differences of same testing corpus under different tagging layers. The performance of the joint model is almost unaffected by the mixed corpus, so the experiment can verify the effectiveness of the integrated tagging method of word segmentation, POS tagging and sentence segmentation.

The experiment selects BiLSTM-CRF as model, mixed corpus as training corpus, and 128-word vector dimensions. The experimental results are shown in Table 7.

Tagging layers		*Tso Chuan*			*Brush Talks from Dream Brook*			*Fantastic Tales by Ji Xiaolan*			*Documents of History of Qing Dynasty*		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
only sentence segmentation		83.6	79.5	81.5	69.0	64.4	66.6	68.1	68.7	68.4	86.8	83.9	85.3
only word segmentation		88.8	91.4	90.0	87.4	85.8	86.6	85.8	87.1	86.4	81.2	77.2	79.2
POS	word segmentation	88.9	91.2	90.0	86.9	86.2	86.6	85.5	86.8	86.1	82.1	77.4	**79.7**
	POS tagging	79.2	81.2	80.2	67.6	65.6	**66.6**	72.2	73.2	72.7	71.8	67.7	69.7
integration	sentence segmentation	86.5	81.9	**84.1**	72.0	71.1	**71.5**	73.7	73.0	**73.3**	85.2	88.8	**86.9**
	word segmentation	89.4	91.1	**90.2**	87.6	85.9	**86.8**	86.3	87.0	**86.6**	81.7	77.0	79.3
	POS tagging	80.1	81.7	**80.9**	67.4	65.4	66.4	72.5	72.9	**72.7**	72.8	68.6	**70.6**

Table 7: Experimental results of BiLSTM-CRF model based on mixed corpus applying to various corpus under different tagging layers

After comparing model’s tagging results of each testing set under different tagging layers, there are 4 conclusions:

(1) By observing the F1-score of each testing set in the same tagging layer, it is found that taking mixed corpus as training set, tagging results of the model applying to various testing corpus are not balanced, which are similar to experiment 3’s result. By comparing the results under the layer that integrates sentence segmentation and lexical analysis with experiment 3, we found that *Brush Talks from Dream Brook*'s performance in sentence segmentation, word segmentation and POS tagging tasks are 0.7, 1.0, 0.5 percentage points higher respectively; *Fantastic Tales by Ji Xiaolan*'s performance in sentence segmentation and POS tagging tasks are 2.3, 0.7 percentage points higher respectively; *Tso Chuan* declines slightly in all tasks. This result indicates that the integration model based on mixed corpus has learnt some homogeneity features of each corpus, which improves some testing sets’ tagging performances. However, in the meantime, the differences among corpus interferes with the comprehensive judgment of the model, resulting in some testing sets’ performance degradation. Therefore, the generalization ability of the integrated tagging model applying to different ages’ texts needs to be improved.

(2) By observing the F1-score of each testing set’s word segmentation task under different tagging layers, the layer

that integrates sentence segmentation and lexical analysis performances best in its entirely. Regardless of which testing set, the F1-score of the tagging layer that only segments word is lower than the integrated layer, which means that integrated tagging method of sentence segmentation and lexical analysis can improve word segmentation task in ancient Chinese.

(3) By observing the F1-score of each testing set's sentence segmentation task under different tagging layers, the layer that integrates sentence segmentation and lexical analysis performances best in its entirely, which shows that integrated tagging method can improve sentence segmentation task in ancient Chinese. Taking *Tso Chuan* as example, the F1-score of sentence segmentation under integrated tagging layer is 2.6 percentage higher than the layer only segment sentence. Similar improvement happens in other testing sets, reflecting that in automatic sentence segmentation task of ancient Chinese, integration of sentence segmentation and lexical analysis is better than step-by-step tagging method.

(4) Comparing the layer of integrated tagging and the layer of POS tagging, we can find that the F1-score of integrated tagging in most testing sets is higher than POS tagging layer. Taking *Tso Chuan* as example, the performance of word segmentation and part-of-speech tagging under integrated tagging layer is 0.2 and 0.7 percentage higher than the POS tagging layer respectively. This result verifies that the integration of sentence segmentation and lexical analysis performances better in word segmentation task and POS tagging task than those methods without adding information of sentence break.

A comprehensive analysis based on (2), (3), (4) can find that the sentence segmentation, word segmentation, and POS tagging tasks have improvement because of the integrated annotation system, and the promotion(F1-score) is not limited to one kinds of testing set. The concrete conditions are shown in Table 8.

Tagging tasks	*Tso Chuan*	*Brush Talks from Dream Brook*	*Fantastic Tales by Ji Xiaolan*	*Documents of History of Qing Dynasty*
sentence segmentation	+2.6	+4.9	+4.9	+1.6
word segmentation	+0.2	+0.2	+0.2	+0.1
POS tagging	+0.7	-0.2	+0	+0.9

Table 8: The promotion of F-score in each task after using the integrated annotation system

Although the integrated tagging method has limit in task promotion, the experiment proves the feasibility of it. It can avoid multi-level spread of tagging errors in single task. For example, if performing tasks step-by-step, we need segment sentence first, and then perform word segmentation task and POS tagging task, which will cause erroneous multi-level accumulation, and the whole performance is not as good as the integrated method. What's more, the tagging method of integrating sentence segmentation and lexical analysis can greatly improve the efficiency of processing words and sentences in ancient Chinese.

7. Conclusion

This paper designs and implements the annotation systerm of integrating sentence segmentation and lexical analysis of ancient Chinese. Based on BiLSTM-CRF neural network model, we verify the intergrated tagging model's generalization ability on different ages' texts, as well as the model's effects on sentence segmentation, word segmentation and part of speech tagging of ancient Chinese under different tagging layers on four different historical testing sets, including *Tso Chuan*, *Brush Talks from Dream Brook*, *Fantastic Tales by Ji Xiaolan* and *Documents of History of Qing Dynasty*. The results appeal that the integrated tagging method performs better among tasks of sentence segmentation, word segmentation and POS tagging. The F1-score of sentence segmentation reached 78.95, with an average increase of 3.5%; the F1-score of word segmentation reached 85.73%, with an average increase of 0.18%; and the F1-score of part-of-speech tagging reached 72.65, with an average increase of 0.35%.

Future research will expand the scale of corpus and improve the model. Focusing on the design of deep learning model in the context of large-scale cross era corpus, the model will include *attention* system and transfer learning method to explore the adaptability of model to different times' texts. Finally, we will develop an integrated analysis system of ancient Chinese with better performance across the ages and styles.

8. Acknowledgements

Supported by the project for Jiangsu Higher Institutions& Excellent Innovative Team for Philosophy and Social Sciences.

A Project Funded by the Priority Academic Program Deve lopment of Jiangsu Higher Education Institutions.

National Language Committee Project of China (YB135-61).

Researches of Building Syntactic Chinese-English Parallel Corpus and Humanities Computing Based on Ancient Classics Index(NO. 71673143) national natural science foundation of China.

9. Bibliographical References

Chen J, Weihua L I, Chen J I, et al. Bi-directional Long Short-term Memory Neural Networks for Chinese Word Segmentation[J]. Journal of Chinese Information Processing, 2018, 32(2):29-37.

HAN X, WANG H, Zhang S, et al. Sentence Segmentation for Classical Chinese Based on LSTM with Radical Embedding[J]. The Journal of China Universities of Posts and Telecommunications, 2019, 26(2):1-8.

Hochreiter S, Schmidhuber, Jürgen. Long Short-Tern Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

Kaixu Z, Yunqing X, Hang Y. CRF-based approach to sentence segmentation and punctuation for ancient

Chinese prose[J]. Journal of Tsinghua University (Science and Technology), 2009(10):1733-1736.

Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv: 1301.3781, 2013.

Min S, Bin LI, Xiaohe C. CRF Based Research on a Unified Approach to Word Segmentation and POS Tagging for Pre-Qin Chinese[J]. Journal of Chinese Information Processing, 2010, 24(2):39-46.

Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088):533-536.

Strubell E, Verga P, Belanger D, et al. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions[C] // Proceeding of the 2017 Conference on Empirical Methods in Natural Language.

Wang B, Shi X, Tan Z, et al. A sentence segmentation method for ancient Chinese texts based on NNLM// Proceedings of CLSM. Singapore, 2016: 387-396.

XUE N. Chinese word segmentation as character tagging [J]. Computational Linguistics and Chinese Language Processing, 2003, 8(1): 29-48.

Yao Y, Huang Z. Bi-directional LSTM Recurrent Neural Network for Chinese Word Segmentation[C]// Conference on Empirical Methods in Natural Language Processing. 2016:1197-1206.

Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.

Yun-Tian F, Hong-Jun Z, Wen-Ning H, et al. Named Entity Recognition Based on Deep Belief Net[J]. Computer Science, 2016, 43(4):224-230.

Zheng X, Chen H, Xu T. Deep Learning for Chinese Word Segmentation and POS Tagging[C] // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013:647-657.

10. Language Resource References

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