Term Translation Extraction from Historical Classics Using Modern Chinese Explanation

Xiaoting Wu [0000-0002-3630-0449], Hanyu Zhao [0000-0002-8436-6397], Chao Che✉

Dalian University, Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, Dalian, 116622, China

wuxiaoting2017@163.com
hanyuzhao7@163.com chechao101@163.com

Abstract. Extracting term translation pairs is of great help for Chinese historical classics translation since term translation is the most time-consuming and challenging part in the translation of historical classics. However, it is tough to recognize the terms directly from ancient Chinese due to the flexible syntactic of ancient Chinese and the word segmentation errors of ancient Chinese will lead to more errors in term translation extraction. Considering most of the terms in ancient Chinese are still reserved in modern Chinese and the terms in modern Chinese are more easily to be identified, we propose a term translation extracting method using multi-features based on character-based model to extract historical term translation pairs from modern Chinese-English corpora instead of ancient Chinese-English corpora. Specifically, we first employ character-based BiLSTM-CRF model to identify historical terms in modern Chinese without word segmentation, which avoids word segmentation error spreading to the term alignment. Then we extract English terms according to initial capitalization rules. At last, we align the English and Chinese terms based on co-occurrence frequency and transliteration feature. The experiment on ***Shiji*** demonstrates that the performance of the proposed method is far superior to the traditional method, which confirms the effectiveness of using modern Chinese as a substitute.

Keywords: BiLSTM-CRF, Co-occurrence frequency, Transliteration features, Term translation extraction.

1 Introduction

Translating outstanding Chinese classics into English is an essential way for Chinese culture promotion. However, at present, only about 0.2% of classical books in China are translated into foreign languages [1]. Speeding up the translation of classics by machine translation is imperative. However, the existing machine translation trained for modern Chinese cannot generate a good translation for historical classics due to the enormous grammatical difference between ancient Chinese and modern Chinese, which can be demonstrated by the example in Table 1. As shown in Table 1, two state-of-the-art machine translation systems, Google and Baidu both give the wrong translation for a sentence from historical book ***Shiji***, which is far from correct translation. Besides, it is tough to construct a machine translation system for ancient Chinese because of the lack of ancient Chinese-English parallel corpora. Now it is feasible to carry out targeted research on the most challenging part of historical classics translation, namely, term translation. This paper extracts the term translation pairs automatically from the bilingual corpus and constructs the term translation dictionary to provide a reference for translators.

Table 1. The translation of the state-of-the-art machine translation systems for an ancient Chinese sentence

Ancient Chinese	固问，语三日，缪公大说，授之国政，号曰五羖大夫。
Google translation	Asked the question, on the 3rd day, Gong Gongda said that he granted the government of the country and the doctor of the 5th class.
Baidu translation	When asked, three days later, Miao Gong said that he gave the national government the number five doctors.
Correct translation	He persisted in his questioning, and they talked for three days. Duke Mu was overjoyed and wanted to hand over the governing of the state to him, entitling him Lord Five Ram Skins.

Accurately identifying terms is the prerequisite for term translation extraction. Nevertheless, it is challenging to extract terms from ancient Chinese directly for three reasons. First, ancient Chinese often omits the context words around the term and uses more flexible grammar; Second, the tagged corpus for ancient Chinese is very limited; Third, there are no efficient algorithms for ancient Chinese word segmentation. Compared to ancient Chinese, term recognition of modern Chinese is more efficient because modern Chinese has a larger scale of tagged corpus, more sophisticated term recognition algorithms [2, 3], and more normative form of language expression. Since most of the terms in ancient Chinese are still preserved in modern Chinese, it is feasible to use the modern Chinese translation of the ancient Chinese to realize the extraction of translated pairs of ancient terms. Moreover, the spread of word segmentation errors can significantly reduce the performance of entity recognition for the traditional Chinese term recognition methods based on words. To this end, this paper proposes a term translation extraction method using character-based sequence model to recognize terms from corresponding modern Chinese explanation instead of ancient Chinese historical books. Specifically, we use the character-based BiLSTM-CRF model [4-6] to identify ancient Chinese terms and extract English terms by the rule of the first letter capitalization. Then extract the Chinese-English term translation pairs according to the co-concurrence frequency [7] and transliteration feature. The framework of term translation extraction is illustrated in Figure 1.

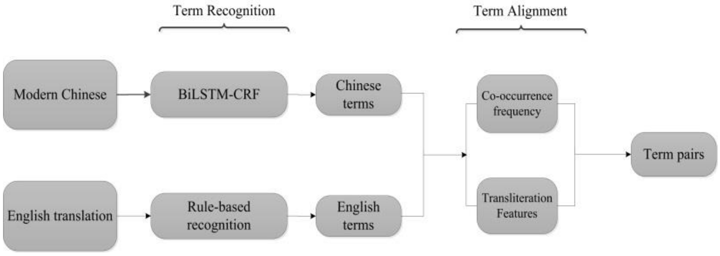

Fig. 1. The framework of term translation extraction method using the character-based model to identify term from modern Chinese instead of ancient Chinese

2 Methodology

According to the framework shown in Fig.1, this section gives the detailed introduction of each step of the term translation extraction method.

2.1 Recognition of Chinese Terms

Terms in historical classics refer to names, places, official titles and posthumous titles, which are similar to the named entities in modern Chinese. Therefore, extracting terms can be viewed as a name entity recognition(NER) task. Traditional NER methods can be classified into rule based method [8] and statistical machine learning method. The commonly used machine learning methods include Maximum Entropy (ME) [9], SVM [10], CRF [11, 12] and HMM [13]. In recent years, Deep Learning NER method [14] has received extensive attention in the field of natural language processing. Compared to rules-based and statistical machine translation methods, Deep Learning methods have the advantage of generalization ability and less reliance on artificial features.

We employ a character-based BiLSTM-CRF hybrid model to conduct ancient term recognition. The LSTM [15] model has a certain memory function which can effectively solve the problem that the traditional recurrent neural network cannot handle long-distance dependence well. In the sequence labeling task, it is usually necessary to consider the context information at the same time. The unidirectional LSTM only considers the past features, so that the comprehensive feature information cannot be obtained. Therefore, a bidirectional LSTM model was adopted to solve the problem in this paper. The forward LSTM model can record the history information of the sequence and the reverse LSTM model can record the future information. Finally, the outputs of the two models are spliced and used as the final output of the hidden layer. CRF is the most common machine learning model in NER. The objective function of CRF not only considers the input feature, but also takes the label transfer feature into account. In other words, the current prediction label is related to the current input feature and the previous prediction label, and there is a strong interdependence between the prediction label sequences. Adding a CRF layer after the LSTM layer can avoid the occurrence of a label "O" followed by a label "I" when using the " BIO " label strategy [16] for NER. So, the BiLSTM-CRF model can use both the past and the future features learned in the bidirectional LSTM layer, as well as the annotation information learned at the sentence level in the CRF layer.

Table 2. Example of annotation method

Original sequence	五	大	夫	吕	礼	出	走	逃	亡	到	魏	国	。
Tagged sequence	B	I	I	B	I	O	O	O	O	O	B	I	O

We use the "BIO" strategy to annotate the corpus. The labeling method is shown in Table 2. "B" denotes the first word of the term, "I" denotes the non-initial word of the term, and "O" denotes the non-term.

The model structure of Chinese term recognition used in this paper is shown in Fig. 2.

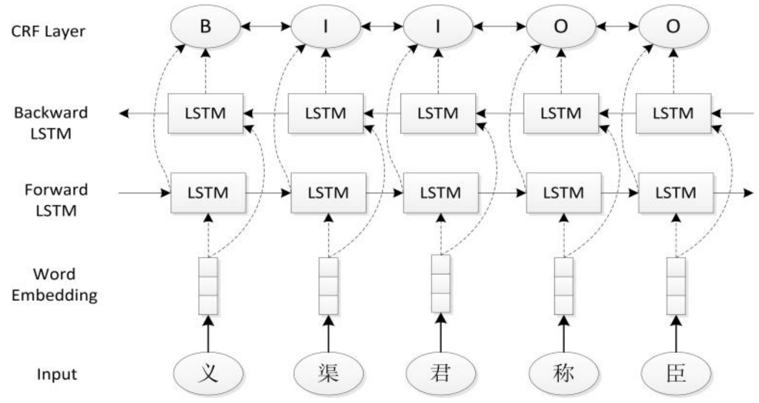

Fig. 2. The structure of the Chinese term recognition model

Given a Chinese sentence, $x = (x_1, x_2, \cdots, x_n)$ indicates the corresponding input word sequence, each layer of BiLSTM-CRF model works as follows:

1. The first layer of the model is the embedding layer, which maps each Chinese character in the sentence into a character vector.
2. The second layer of the model is a bidirectional LSTM layer which automatically extracts sentence features. The vector of each character in the sentence is used as the input for each time step of the bidirectional LSTM. Then, the hidden states output by the forward LSTM $\left(\overrightarrow{h_1}, \overrightarrow{h_2}, \cdots, \overrightarrow{h_n}\right)$ and the reverse LSTM $\left(\overleftarrow{h_1}, \overleftarrow{h_2}, \cdots, \overleftarrow{h_n}\right)$

are bit-spliced as $h_t = [\vec{h}_t; \overleftarrow{h}_t] \in R^m$ to obtain a complete hidden state quence $(h_1, h_2, \cdots, h_n) \in R^{n \times m}$. The probability output matrix of LSTM is defined as P_{n*k}, where k is the number of output labels, $P_{i,j}$ refers to the probability that the i-th word is marked as the j-th tag.

3. The third layer of the model is the CRF layer, where sentence-level sequence annotation is performed. For the tag sequence to be predicted: $y = (y_1, y_2, \cdots, y_n)$, the score of the model for the tag of the sentence x equaling to y is defined as follows:

$$s(x,y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=0}^{n} p_{i,y_i} \tag{1}$$

Where A is the state transition matrix and $A_{i,j}$ represents the probability of the transition from the i-th label to the j-th label. y_0 and y_n are the labels added to the position at the beginning and end of the sentence, respectively. So, A is a square matrix of size k+2.

Softmax function is employed to get the normalized probability:

$$P(y \mid x) = \frac{\exp(score(x,y))}{\sum_{y'} \exp(score(x,y'))} \tag{2}$$

Model is trained by maximizing the logarithmic probability of correct label sequence:

$$\log P(y^x \mid x) = score(x, y^x) - \log(\sum_{y'} \exp(score(x, y'))) \tag{3}$$

The model uses the viterbi algorithm of dynamic programming to obtain the best output label sequence during the prediction process:

$$y^* = \arg\max_{y'} score(x, y') \tag{4}$$

2.2 Recognition of English Terms

This paper uses ***Record of the Grand Historian of China*** [17] as the English translation, in which all the English terms are capitalized. Consequently, we use the initial capitalization rules to identify English terms. However, the capitalization extraction rule has two problems: (1) the first word of the sentence is extracted as a wrong term and (2) some articles and conjunctions in terms will be missed, such as "the " and "of", which are not capitalized in terms. For the first problem, we do not treat it as a term when the extracted term is at the beginning of the sentence and contains only one word of following part of speech: numeral, preposition, adverb, conjunction, etc. For the second problem, if "the" is followed by a capital word, or "of" is sandwiched between two capital words, they are added to the extracted terms.

2.3 Term alignment

According to the alignment process, the alignment method based on bilingual parallel corpus can be divided into two categories: (1) the symmetric method, which identifies the terms in two languages, respectively, then the alignment model is used to align the terms in the two sides; (2) the asymmetric method, which recognizes the terms in one language, then find its corresponding translation in another language. The recognition methods used in this paper have good performance on Chinese terms and English terms. Therefore, we use the symmetric method to align terms with co-concurrence frequency and transliteration feature.

Co-occurrence frequency. To avoid the performance decrease of term alignment caused by word segmentation errors, we adopt a character-based Chinese term recognition method. However, this method can identify many non-terms, which can negatively affect the term alignment and increase the difficulty of alignment. Considering the co-occurrence frequency of terms is helpful for identifying the translation of terms accurately. If a term pair appears more frequently in all the pairs of terms which are related to English term e, then it is more likely to be a correct term pair.

For an English term e, the co-occurrence frequency of term pairs is defined as:

$$F(c_i \mid e) = \frac{N(c_i, e)}{N_e} \tag{5}$$

Where $N(c_i, e)$ denotes the times that the term pair co-occurs, and N_e denotes all term pairs that contain the English term e.

Transliteration feature. Transliteration is often used in the English translation of historical books. According to two kinds of transliteration of historical classics, we use the method in [18] for reference and use the proportion of the transliteration words in the English terms as transliteration feature values. The transliteration function can be defined as follows:

$$H(c \mid e) = \frac{N_{pinyin}(c, e) + N_{title}(c)}{len(e)} \tag{6}$$

Wherein, Len(e) indicates the number of the real words in English terms, and the prepositions such as "the", "of", "in", "on" are not included in the counting. The word containing "-" is counted as two words. $N_{pinyin}(c, e)$ denotes the number of Chinese phonetic alphabet in the English term e corresponding to the Chinese term c; $N_{title}(c)$ indicates the number of characters of fixed title in the Chinese term c. We construct a fixed title list manually. The number of fixed title can be determined by querying the list.

Term alignment. For each candidate Chinese term of the English term e, we calculate the co-occurrence frequency F and the transliteration feature value H, which are added to obtain the probability of alignment. Finally, the candidate Chinese term with the largest alignment probability is selected as the translation of the term e.

2.4 Term translation extracting method

Given a bilingual corpus, the algorithm for extracting Chinese and English term translation pairs is shown in Fig. 3.

```
Input: Chinese-English corpus CE
(1) For every parallel sentence pair $(c_i, e_i), i \in (1, n)$, extract Chinese
    terms $A_i = \{t_1^c, t_2^c, \cdots, t_m^c\}$ and English terms $B_i = \{t_1^e, t_2^e, \cdots, t_s^e\}$
    from Chinese and English sentence, respectively.
(2) For every English term $t_i^e \in B_i$, calculate the co-occurrence frequency $F(t_j^c|t_i^e)$, $t_j^c \in A_i$ .
(3) For every English term $t_i^e \in B_i$, calculate the transliteration feature
    value $H(t_j^c|t_i^e)$, $t_j^c \in A_i$ .
(4) For each candidate term pair$(t_j^c, t_i^e)$, calculate the alignment probability: $P(t_j^c|t_i^e) = F(t_j^c|t_i^e) + H(t_j^c|t_i^e)$, and select the term pair
    with the greatest probability of alignment $P$ for $t_i^e$ as the final result
    to add to the set $T_{ce}$.
Output: Chinese - English term pairs $T_{ce}$
```

Fig. 3. Term translation extraction algorithm

3 Experiment

3.1 Experimental Setup

In the experiment, we used the parallel corpora composed of modern Chinese and English translation of ***Shiji*** to extract historical term pairs. The modern Chinese we used is ***The Chronicle of the Vernacular History*** [19], and the English translation is the 1961 version of ***Record of the Grand Historian of China*** which is written by Burton Watson. Based on the existing ancient Chinese-English bilingual corpus, we find the corresponding modern Chinese for each ancient Chinese sentence to construct the modern Chinese-English bilingual parallel corpora. The training set contains 5060 sentence pairs which are selected from ***Annals of Qin, the Basic Annals of the First Emperor of the Qin, the Basic Annals of Hsiang Yu, the basic Annals of Emperor Kao-tsu and the Basic Annals of Empress Li***. 1085 sentence pairs from ***the Hereditary House of the Marquis of Liu*** *and* ***the Tiereditary House of Prime Minister***

Chen are used as a test set. And 287 term translation pairs are manually extracted from test set as standard answer.

3.2 Evaluation Metrics

The experiment results are evaluated by precision (P), recall (R), and F1 measure. Suppose N_{gold} is the number of term pairs in our corpus, $\mathrm{N}_{extracted}$ is the number of the term pairs that extracted by our method, and $N_{correct}$ is the number of correct term pairs extracted by our method. P, R, F1 are defined as follows, respectively.

$$P = \frac{N_{correct}}{N_{extracted}} \times 100\% \tag{7}$$

$$R = \frac{N_{correct}}{N_{gold}} \times 100\% \tag{8}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

3.3 Experimental Result and Analysis

Feasibility Analysis. To verify the feasibility of using modern Chinese to replace ancient Chinese, we conducted term translation extracting method based on BiLSTM-CRF to extract the term pairs from the ancient Chinese-English corpus and the modern Chinese-English corpus, respectively. The results comparison is shown in Table 3.

Table 3. Result comparison of the term translation extraction method using different corpora

corpus	P	R	F1
ancient Chinese-English	67.4%	53.5%	59.7%
modern Chinese-English	**80.8%**	**79.4%**	**80.1%**

From Table 3, it can be seen that the method of extracting terms from modern Chinese significantly improved the P, R, and F1 values compared with the method of directly extracting terms from ancient Chinese. The main reason is that the ancient Chinese often adopts complex and flexible sentence structures such as passive, inversion, and omit structure. When the corpus is relatively small, it is tough to learn accurate and comprehensive features using the BiLSTM-CRF model. Therefore, the performance of the term translation extraction method using ancient Chinese is poor. For example, the sentence "四人至，客建成侯所。"can be translated into modern Chinese as "四个人来了，就住在建成侯的府第中为客。"(When the four arrived, they were entertained as guests at the house of LüZe). The sentence "客建成侯所" reflects a more common structure of the object ahead of the predicate in ancient Chinese. In addition, due to the richness of lexical meaning of words in ancient Chinese, a noun "客" expresses the meaning of "to be a guest in …". It is indeed difficult for LSTM to learn similar features in ancient Chinese, so the term “建成侯” was not identified in this sentence. In modern Chinese, this complex sentence structure has been transformed into a simpler and more regular structure of “subject-predicate-object”, which is more conducive to the study of characteristics. Therefore, “建成侯” can be successfully identified in modern Chinese corpus.

According to the statistics of experimental corpus, about 96% of the ancient Chinese terms are completely preserved in modern Chinese, which means that the majority of term translation pairs can be extracted by this method. However, for some ancient Chinese terms that have not been completely preserved in modern Chinese, such as interchangeable words, we cannot obtain their translations. For example, the modern Chinese translation “宁昌” corresponding to ancient term “甯昌”, therefore, we can not extract the enlish translation of “甯昌” through the method based on modern Chinese explanation.

The effect of term recognition on term alignment. The quality of the term recognition directly affects the effect of term alignment. So, we compared BiLSTM-CRF with CRF and LSTM-CRF to explore the effect of different term recognition methods on the term pair extraction. The results of the experiment are shown in Table 4.

Table 4. Result comparison of different term recognition methods

Method	Term Recognition			Term Alignment		
	P	R	F1	P	R	F1
CRF	77.3%	70.3%	73.6%	67.3%	61.2%	64.1%
LSTM-CRF	84.4%	79.4%	81.8%	72.5%	68.5%	70.4%
BiLSTM-CRF	**89.3%**	**87.8%**	**88.5%**	**80.8%**	**79.4%**	**80.1%**

Term recognition is the basis for the term alignment, and higher term recognition recall rate ensures that as many term as possible are identified, which lays the foundation for the term alignment. From Table 4, we can see that the higher the recall rate of the term recognition, the better the result of term alignment.

Comparison with traditional term alignment methods. In this paper, the traditional IBM model4 alignment model was used as the baseline. The IBM model4 method employed Jieba[1] to perform word segmentation for modern Chinese and then aligned the term between modern Chinese and English by GIZA++[2]. The results of the experiment are shown in Table 5.

Table 5. Result comparison of different term translation extraction methods

Method	P	R	F1
IBM model4	67.1%	54.2%	60.0%
Our method	**80.8%**	**79.4%**	**80.1%**

[1] https://github.com/fxsjy/jieba

[2] https://codeload.github.com/moses-smt/giza-pp/zip/master

Table 5 shows that the proposed method outperforms the traditional term alignment model IBM Model4 with a large margin. This is due to the following two aspects:

Firstly, traditional alignment methods often have high requirements on word segmentation performance. Obtaining a good word segmentation result is a prerequisite for extracting correct translation terms. Nevertheless, the most used word segmentation methods have a very poor performance for ancient Chinese, which will decrease the performance of term alignment. For example, in the segmentation result of Jieba "留/侯张良/,/他/的/祖先/是/韩国/人", "侯张良" is segmented as a word, which will align"侯张良" with "Zhang Liang" wrongly. On the other hand, our method identifies Chinese terms at the character level, which avoids the problem that the word segmentation errors decrease the accuracy of term recognition.

Secondly, our method can effectively alleviate the problem of low recognition performance of the Chinese term. Because the BiLSTM-CRF model extracted some non-terminological vocabularies when recognizing Chinese terms, such as "小子", "中间" and "五谷", this result in the vast number differences between Chinese and English terms identified in a sentence pair, which makes it more difficult in aligning terms. However, the recall rate of our term recognition model is 89.3%, which guarantees that we can identify most of the terms and lays the foundation for the extraction of term pairs. For Chinese terms and English terms extracted from a sentence pair, we calculate the co-occurrence frequency and transliteration feature values between each Chinese term and each English term to determine the final alignment result, which reduces the effect of non-terminology words on the extraction of term pairs.

4 CONCLUSIONS

In order to solve the difficulty of term translation in the process of translating ancient Chinese classics, this paper proposes a term translation extracting method using multi-features based on BiLSTM-CRF to extract historical term pairs from modern Chinese-English parallel corpora instead of ancient Chinese-English parallel corpora. Our method not only avoids word segmentation error spreading to the term alignment, but also solves the difficulty of extracting historical term translations directly from ancient Chinese corpus.

However, our method can only identify and align terms preserved in modern Chinese. In the future, we will explore the method which can deal with the terms that are not retained in modern Chinese.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. 61402068) and Support Program of Outstanding Young Scholar in Liaoning Universities. (No. LJQ2015004).

References

1. Huang, Z.X.: English Translation of Cultural Classics and Postgraduate Teaching of Translation in Suzhou University. Shanghai Journal of Translators, 1:56-58 (In Chinese) (2007).
2. Wang, B.: Translation Pairs Extraction from Unaligned Chinese-English Bilingual Corpora (In Chinese). Journal of Chinese Information Processing, 14(6): pp. 40-44 (2000).
3. Yang, P., Hou, H.X., Jiang, Y.P., Jian, Shen, Z., D.U.: Chinese-Slavic Mongolian Named Entity Translation Based on Word Alignment (In Chinese). Acta Scientiarum Naturalium Universitatis Pekinensis, 52(1): pp. 148-154 (2016).
4. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, k., Dyer, C.: Neural Architectures for Named Entity Recognition. pp. 260-270 (2016).
5. Zeng, D., Sun, C., Lin, L., Liu, B.: LSTM-CRF for Drug-Named Entity Recognition. Entropy, 19(6) (2017).
6. Dong, C., Zhang, J., Zong, C., Hattori, M., Di, H.: Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition. in International Conference on Computer Processing of Oriental Languages (2016).
7. Li, X., Che, C., Liu, X., Lin, H., Wang, R.: Corpus-based Extraction of Chinese Historical Term Translation Equivalents (2010).
8. Zhou, K.: Research on Named Entity Recognition Based on Rules (In Chinese). Hefei University of Technology (2010).
9. Hai, L.C., Ng, H.T: Named entity recognition: a maximum entropy approach using global information. in International Conference on Computational Linguistics (2002).
10. Li, L., Mao, T., Huang, D., Tang, Y.: Hybrid Models for Chinese Named Entity Recognition. Proceedings of the Fifth Sighan Workshop on Chinese Language Processing, pp. 72-78 (2006).
11. Şeker, G.A., Eryiğit, G.: Extending a CRF-based named entity recognition model for Turkish well formed text and user generated content1. Semantic Web. 8(5): pp. 1-18 (2017).
12. L. Sun, Y. Guo, W. Tang, et al. Enterprise abbreviation prediction based on constitution pattern and conditional random field. Journal of Computer Applications (2016).
13. N.V. Patil, A.S. Patil, B.V. Pawar. HMM based Named Entity Recognition for inflectional language. in International Conference on Computer, Communications and Electronics (2017).
14. Wang, G.Y.: Research of Chinese Named Netity Recognition based on Deep Learning (In Chinese). Beijing University of Technology (2015).
15. Greff, K., Srivastava, R.K., Koutn ́k, J.,Steunebrink, B.R., Schmidhuber, J.: LSTM: A Search Space Odyssey. IEEE Transactions on Neural Networks & Learning Systems. 28(10): pp. 2222-2232 (2017).
16. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. in Conference on Empirical Methods in Natural Language Processing (2008).
17. Watson, B.: Record of the Grand Historian of China[J]. Journal of Asian Studies, 22(2):205 (1961).
18. Che, C., Zheng, X.J.: Sub-Word Based Translation Extraction for Terms in Chinese Historical Classics (in Chinese). Journal of Chinese Information Processing, 30(3): pp. 46-51 (2016).
19. Sixty professors in Taiwan’s 14 institutions. The Chronicle of the Vernacular History (in Chinese). New World Press (2007).