A method of terminology alignment in Historical Books based on the correspondence between Pivot language and position

Xingchen Xie¹, Xiaoting Wu¹, Xiaopeng Wei² and Chao Che¹

¹Key Laboratory of Advanced Design and Intelligent Computing, Dalian University, Dalian, Liaoning, 116622, China
²Corresponding author’s e-mail: chechao@dlu.edu.cn

Abstract. In view of the most difficult problem of terminology translation in the translation of historical books, this paper extracts term translation pairs from bilingual parallel corpus as a reference to improve the quality and efficiency of the translation of historical books. However, the small size of the corpus of historical books leads to poor word alignment, which brings a large number of errors to term alignment. At the same time, the special grammar of ancient Chinese will also affect term alignment. In order to solve the above two problems, this paper proposes a method of term alignment of historical books based on pivot language and position correspondence. This method introduces modern Chinese as the pivot language, makes use of the position correspondence between modern Chinese and English terms in sentences, and uses the sliding window method to generate candidate term pairs, which avoids the influence of word alignment errors on term alignment. In addition, if we first align to modern Chinese, the ancient Chinese terminology will complete the omitted characters in modern Chinese and replace the false characters, so as to avoid the influence of the special usage in ancient Chinese on the term alignment. The experimental results in Historical Records show that the accuracy of this method is slightly better than that of the method based on word alignment, which verifies the effectiveness of this method without word alignment.

1. Introduction

Reading ancient books is one of the important means to understand a history of civilization and a kind of culture. However, the language barrier also increases the difficulty for the spread of Chinese culture. As a consequence, the English translation of classical books is of great importance. However, only a small part of the vast number of books in our country have been translated, and according to statistics, only about 2/1000 of them have been translated into foreign languages [1]. There are many reasons for this phenomenon, such as the richness of the connotation of Chinese classics, the "time difference" and "language gap" of cultural exchanges between China and the West, and so on, which make the translation task full of challenges. One of the most time-consuming and challenging tasks in the translation of classical books is terminology translation, which directly affects the quality of translation. As a consequence, this paper hopes to extract term translation pairs from bilingual parallel corpus as a reference for the translation of historical books, so as to improve the efficiency and accuracy of the translation of historical books.

The terms in historical books mainly refer to "official title, name, title, etiquette, system, custom name" and so on. It is similar to named entities in natural language processing, so in this paper, term alignment is approximately regarded as named entity alignment. From the perspective of the corpus used, there are three main methods of named entity alignment: (1) obtaining named entity translation
pairs from web pages [2-4]; (2) obtaining named entity translation pairs from bilingual parallel corpus [5-8]; and (3) obtaining named entity translation pairs from comparable corpus [9-11]. Historical terms seldom appear in web pages and comparable corpus. This paper mainly extracts term translation pairs from bilingual parallel corpus. The early term alignment of historical books uses a heuristic method based on mutual information [12]. Since then, in view of the lack of ancient Chinese word segmentation algorithm, Che Chao and Zheng Xiaojun counted the words that frequently appeared together as sub-words and used sub-words to align terms after word segmentation of classical books [13]. Wu et al. in view of the characteristics that most of the ancient Chinese terms are retained in the modern Chinese interpretation, use modern Chinese interpretation instead of ancient Chinese for term alignment, avoiding the influence of the special grammar of ancient Chinese on term alignment [14]. The above term alignment methods are all based on the results of word alignment [15-16], but at present, the performance of the word alignment model is not satisfactory, especially in the parallel corpus of historical books with small corpus. A large number of errors are introduced into term alignment. In addition to the errors caused by word alignment, the special grammatical structure of ancient Chinese will also cause errors in the alignment of historical terms. For example, words in ancient Chinese often use abbreviations, but English uses full names to express, which creates contradictions.

In order to avoid the influence of word alignment and ancient Chinese special usage on term alignment, this paper proposes a term alignment method based on pivot language and position relationship. This method uses modern Chinese as the pivot language, first uses co-occurrence characters to align ancient Chinese terms to modern Chinese terms, and then uses the sliding window method based on position correspondence to align modern Chinese to English translation. When aligning modern Chinese and English terms, we first use the position correspondence in the sentence to find the candidate term pairs in the sliding window, and then use the SVM model which combines the co-occurrence frequency and transliteration features to select the correct term translation pairs. The method based on pivot language is to introduce a third language with parallel corpus between source language and target language as pivot to expand corpus resources [17-19], which is used to solve the problem of corpus shortage in machine translation of low resource languages. In this paper, modern Chinese is used as the pivot language. In addition to expanding the corpus, the main purpose is the following two points: (1) The difference between the grammatical rules of ancient Chinese and modern languages makes it difficult to correspond to ancient Chinese and English. However, ancient Chinese and modern Chinese do not have such problems, so this article adopts the pivot language method to avoid the occurrence of errors. In this paper, by using the pivot language to align ancient Chinese to modern Chinese, the correspondence in this position can be used to select candidate term pairs, which avoids the error of word alignment in term alignment. (2) aligning ancient Chinese terms to modern Chinese terms can supplement the full name of terms, replace the common false words in ancient Chinese, and reduce the alignment errors caused by abbreviations and common false words.

2. Terminology alignment method based on Pivot language

2.1. Term alignment frame

This paper takes modern Chinese as a bridge to realize the alignment between ancient Chinese terms and corresponding English terms. In the English translation of historical books, the vast majority of terms have the characteristics of capitalization, so the terms in English are more convenient to identify, while the term recognition in ancient Chinese requires manual tagging corpus, which is more difficult and the accuracy is not high. Therefore, this paper first uses rules to identify terms from English corpus, uses positional correspondence to obtain translation pairs of English-modern Chinese candidate terms, and uses SVM model combined with co-occurrence frequency and transliteration features to find correct term pairs, and then uses co-occurrence features to realize term pair extraction between ancient Chinese and modern Chinese. Finally, the corresponding relationship between ancient Chinese terms and English terms is established with modern Chinese as the pivot language. The alignment frame is shown in figure 1.
Next, we will introduce in detail the extraction rules of English terms and the process of term alignment from ancient Chinese to modern Chinese to English.

2.2. English term recognition
This paper mainly uses the initial capitalization rules to identify English terms. Most of the English terms in the section translation of Historical Records of BurtonWatson are capitalized. However, if you simply extract all the words with uppercase initials, there will be the following two problems: (1) the capitalization of the first letter of a sentence is extracted incorrectly as a term. (2) sometimes not every word in English terms is capitalized, such as "the", "of" and so on. Only words with capitalized initials will be omitted. For the first question, when the extracted term is at the beginning of a sentence or is composed of a single word, judge whether it is a numeral, preposition, adverb, conjunction, etc in order to eliminate the interference caused by the capitalization of the first word of the sentence in the English corpus. For question (2), if "the" is followed by an uppercase word during term extraction, the two words are combined, and if "of" is sandwiched between two capitalized words, "of" should appear in the term.

2.3. The extraction of Modern Chinese-English term pairs
The basic idea of modern Chinese-English term pair extraction: for all the terms in English sentences, the sliding window method based on position correspondence is used to find out all the candidate translations of modern Chinese terms corresponding to the term. Then, according to the SVM model, the candidate terms are screened by combining transliteration features and co-occurrence frequency features, and finally the translation of modern Chinese terms aligned with English terms is determined.

2.3.1. Sliding window method based on position correspondence
Due to the lack of bilingual parallel corpus, the classical IBM model is not effective in the alignment between modern Chinese and English, two non-homologous languages. According to the observation of the corpus, it is found that in a bilingual parallel sentence pair, the order in which the terms appear in the bilingual sentence is highly consistent, and the position in the sentence is roughly the same. Therefore, we can delimit the approximate range of candidate terms on the target side according to the order in which the source terms appear and the position in the sentence, so as to avoid the dependence on the result of word alignment in the traditional alignment mode. The mapping function used in this paper is:

$$\text{Term}_{\text{candidate}} = \frac{w(e)}{l\text{en}(E)} \times \text{len}(C) + \lambda, \lambda \in (-4,4)$$  (1)

Figure 1. Alignment framework of ancient Chinese-English terms based on pivot language
In the formula, $w(e_i)$ denotes the position of the first word of the English term $e_i$ in the sentence $E$ the subscript, $\text{len}(E)$ indicates that the length of the English sentence, $\text{len}(C)$ is the length of the Chinese sentence, and, $\lambda$ is a variable parameter, that is, the size of the context sliding window, which is set to 4 in this paper. The following examples are illustrated in detail.

As shown in figure 2, according to the English term recognition rules, "EmperorGaozu" and "theKingofHuainan" in this sentence can be identified as English terms. Take "EmperorGaozu" as an example, in the sentence, the first subscript of the term is 5. According to formula 1, it can be calculated that the subscript of the candidate Chinese term is $\{1, 2, 3, 4, 5, 6\}$, that is, the candidate terms are: \{"now", "Gaodi", "of", "son", "just", "only"\}.

2.3.2. Characteristic function

After obtaining the pairs of candidate terms, we use co-occurrence frequency and transliteration features to screen all candidate terms and select the Chinese terms corresponding to English terms.

(1) Co-occurrence frequency

Considering the co-occurrence frequency of terms helps to accurately identify the translation of terms: the more frequently a term pair $(c_i, e)$ appears in all the term pairs about $e$, the more likely $(c_i, e)$ is that $(c_i, e)$ is the correct term pair. For an English term $e$, the frequency of co-occurrence of the term to $(c_i, e)$ is defined as:

\[
L(c_i | e) = \frac{N(c_i, e)}{N_e}
\]

(2)

In the formula, $N(c_i, e)$ denotes the number of common occurrences of the term $(c_i, e)$, $N_e$ denotes all pairs of terms that contain the English term $e$.

(2) Transliteration features

Transliteration is often used in the English translation of the names of people, place names and official names in historical books. According to the two transliteration modes of historical books, this paper draws lessons from the reference [13], uses the proportion of transliterated words in English terms as the transliteration eigenvalue, and defines the transliteration function as follows:

\[
H(c | e) = \frac{N_{\text{pinyin}}(c, e) + N_{\text{title}}(c)}{\text{len}(e)}
\]

(3)

Len(e) denotes the number of remaining notional words in English terms after excluding articles and prepositions such as "the", "of", "in" and "on", and words containing the conjunction "." are counted as two words. For example, len ("the Marquis of Wu-fang") = 3; $N_{\text{pinyin}}(c, e)$ indicates that the English term $e$ contains the number of Hanyu pinyin corresponding to the characters in the Chinese term $c$; $N_{\text{title}}(c)$ indicates that the Chinese term $c$ contains the number of characters with fixed appellations,
and whether it contains fixed appellations is determined by querying the list of manually constructed fixed appellations.

2.4. Extraction of Ancient Chinese-Modern Chinese term pairs

Although ancient Chinese and modern Chinese are not strictly speaking a language, but as a cognate language, the use of the same character similarity makes it easier to align terms between them. In addition, modern Chinese originates from ancient Chinese, most of the terms in ancient Chinese have been effectively retained in the process of language evolution, and a few terms (such as personal names and official names) have only been supplemented and perfected on the original basis. The supplement and perfection of modern Chinese to ancient Chinese terms are mainly divided into two categories: one is the supplement to the surnames of characters. In Chinese, the first name is composed of the first name and the surname. In ancient Chinese, the surname of the character is often omitted and only the first name is used to refer to the person. For example, in the corpus, "Liang" is often used to refer to "Zhang Liang" and the surname "Zhang" is omitted. The second is the perfection of place names and titles of characters. For example, in ancient Chinese, "Gui" refers to "Guishan". In modern Chinese, in order to make the expression of place names more accurate, "Guishan" is all translated into "Gushan". The omission of titles is also common in ancient Chinese corpus, as shown in Table 1. In this sentence, specific titles are omitted after "Qiwei", "Chu Xuan", "Wei Hui", "Yan mourning" and "Han Ai". The complete expression should be the translation in modern Chinese, that is, "King of Qi Wei", "King of Chu Xuan", "King Wei Hui", "Yan mourning King" and "Han Ai Hou". Based on the above two ellipsis phenomena, ancient Chinese terms are usually a subset of modern Chinese term strings. Therefore, this paper realizes the extraction of ancient Chinese-modern Chinese terminology translation pairs according to the co-occurrence of term characters in the two languages.

Table 1. Examples of omission of titles in ancient Chinese

| Ancient Chinese | Modern Chinese translation | English translation                                      |
|-----------------|----------------------------|--------------------------------------------------------|
| 与齐威、楚宣、魏惠、燕悼、韩哀、赵成侯并。 | 秦孝公与齐威王、楚宣王、魏惠王、燕悼王、韩哀侯、赵成侯并称。 | The ruler King Wei of Qi, King Xuan of Chu, King Hui of Wei, Duke Dao of Yan, Duke Ai of Han, Duke Cheng of Zhao being ranged side by side. |

3. Experiment

3.1. Experimental setup

This paper uses the ancient Chinese-English bilingual parallel corpus (among which, the ancient Chinese uses Shi ji and the English translation uses the 1961 version of Record of the Grand Historian of China of Burton Watson) and the modern Chinese-English bilingual parallel corpus (among them, the vernacular Historical Records is used in modern Chinese). The parallel corpus of "Qin Benji", "Qin Shi huang Benji", "Xiang Yu Benji", "Gao zu Benji" and "Empress Dowager Lu Benji" in Historical Records is selected as the training set, and the 1085 sentences of "Liuhou Family" are selected as the test set.

3.2. Evaluation index

In this experiment, accuracy P, recall R and F1 value are used as evaluation indicators of term alignment, which are defined as follows:

\[ P = \frac{\text{The term logarithm of the correct alignment of the model}}{\text{The term logarithm of model alignment}} \times 100\% \]  (4)

\[ R = \frac{\text{The term logarithm of the correct alignment of the model}}{\text{logarithm of actual terms in the corpus}} \times 100\% \]  (5)


\[ F1 = \frac{2 \times P \times R}{P + R} \quad (6) \]

3.3. Experimental results and analysis

3.3.1. The effectiveness of pivot language

In the experiment, the term alignment method proposed in this paper is used to realize the ancient Chinese-modern Chinese-English term alignment. Among them, both modern Chinese and ancient Chinese use stuttering participle 1. At the same time, with reference to the ancient Chinese and English direct alignment model, the two methods are compared in terms of accuracy, recall rate and F1 value, and the results are shown in Table 2.

| Terminology alignment method | P     | R     | F-1    |
|------------------------------|-------|-------|--------|
| Direct alignment             | 84.0% | 63.0% | 72.0%  |
| Methods                      | 93.3% | 72.5% | 81.6%  |

As can be seen from Table 2, the accuracy, recall and F1 value of terminology alignment based on this method are significantly higher than those of direct alignment. The reason is that the term alignment method in this paper can effectively avoid alignment errors caused by ancient Chinese ellipsis. In order to keep the language characteristics of concise and concise words, some words are often omitted in historical books. In Historical Records, the general history of biography, the phenomenon of ellipsis is more common, the ellipsis of information increases the difficulty for the direct alignment of ancient Chinese-English terms, and is prone to misalignment of terms. As shown in figure 3. In the ancient Chinese corpus, the word "TianFen" is used to refer to "Tian Yi". In the process of ancient Chinese-English direct alignment, the lack of information leads to the alignment error, and the corresponding person name "Qi" can not be found. On the other hand, modern Chinese translation will generally complete the omitted part of ancient Chinese and translate the meaning to be expressed completely. Therefore, the difficulty of term alignment between modern Chinese and English is much less than that between ancient Chinese and English, and the accuracy is higher. When the indirect alignment method based on pivot language is adopted, in the process of alignment from ancient Chinese to modern Chinese, it can be recognized that "Tian" and "Tian" refer to the same term, and because "Tian" is completely translated into "Tian" in the corpus of modern Chinese, the alignment of "Tian" and "TianFen" can be realized in the process of alignment between modern Chinese and English. Therefore, based on the pivotal language of modern Chinese, the corresponding relationship between the ancient Chinese term "TianFen" and the English term "ancient Chinese" can be established.

Figure 3. Example of term alignment based on pivot language
3.3.2. Comparison with IBM4 model

We compare this method with the entity alignment model IBMModel42, which is commonly used in traditional machine translation. The experimental results are shown in Table 3.

| Terminology alignment method | P     | R     | F-1   |
|------------------------------|-------|-------|-------|
| IBM Model4                   | 67.1% | 54.2% | 60.0% |
| Methods                      | 93.3% | 72.5% | 81.6% |

As can be seen from Table 3, IBMModel4 is not effective in realizing the terminology alignment between Chinese and English books. The reason is that IBMModel4 is a statistical alignment model, and when the corpus is very small, the effect of word alignment will be very poor, especially for the alignment between modern Chinese and English, which is a non-homologous language, it is difficult to obtain a good alignment effect. As shown in Table 4, in the alignment result given by IBMModel4, the English term "Bolangsha" aligns the "error" in modern Chinese, and the "error" in modern Chinese aligns the "error" in ancient Chinese, thus giving rise to the wrong translation pair of ancient Chinese-English terms "misalignment-Bolangsha". On the other hand, the term alignment method proposed in this paper uses the features of co-occurrence and transliteration, and there will be no alignment errors such as "misalignment-Bolangsha". As shown in Table 5, IBMModel4 does not give any alignment result of "Chunyu Gong" for this sentence corpus, so it is impossible to obtain the term pair "Chunyu Gong-LordChunyu". In the method proposed in this paper, the sliding window is used to obtain the candidate term pair, and the Chinese term "Chunyugong" can be included in the candidate term set of "LordChunyu". Therefore, the term pair can be aligned correctly.

This method is not effective for the term alignment of some inverted sentences, because the coverage of the candidate terms is difficult to include the correct Chinese terms, which directly affects the effect of term alignment. In addition, the result of word segmentation will also affect the alignment of terms. For example, in "Wu Wang Kui, Chu Wang Wu, Zhao Wang Sui, Jiaoxi Wang Kui, Jinan Wang Piguang, disaster Chuan Wang Xian, Jiaodong Wang Xiongqu, send troops to Xixiang." In the process of ancient Chinese word segmentation, both "Piguang" and "Xiongqu" have word segmentation errors, which are divided into "Piguang" and "Xiongqu", resulting in the misalignment of "Piguang-LiuBiguang" and "Xiongqu-LiuXiongqu".

| Ancient Chinese | Modern Chinese | English                                    |
|----------------|----------------|--------------------------------------------|
| 良/与/客/狙/击/秦皇帝/博浪沙/中/, /误中/副车/。 | 张良/与/大力士/在/博浪沙/这个地方/袭击/秦始皇/, /误中/了/副车/。 | He and the assassin lay in wait for him. When the Emperor reached the area of Bolangsha, they made their attack, but mistakenly struck the carriage of his attendants. |

Table 4. Compared with the traditional IBM Model4, example 1

| Ancient Chinese | Modern Chinese | English |  |
|----------------|----------------|---------|--------|
| 五月/, /齐/太仓令淳于公/有/罪/当/刑/, /诏狱/逮捕/他/, /把他/押解到/长安/拘禁/起来/。 | 五月/, /齐国/的/太仓令淳于公/犯了罪/, /应该受刑/, /朝廷/下令/将/狱官/逮捕/他/, /把他/押解到/长安/拘禁/起来/。 | In the fifth month Lord Chunyu, the Chief of the Treasury in the state of Qi, was accused of some crime and sentenced to punishment. Orders were issued to have him bound and escorted under guard to the capital for imprisonment. |

Table 5. Compared with the traditional IBM Model4, example 2
4. Conclusion
In view of the influence of the result of word alignment on term alignment in historical books, this paper proposes a method of term alignment without word alignment. This method takes modern Chinese as the pivot language, establishes the corresponding relationship between modern Chinese and English translation, and then uses the sliding window method based on position correspondence to align terms. The experimental results in Shiji show that the indirect alignment of ancient Chinese-English terms based on the pivot language is significantly higher than the direct alignment between the two languages in terms of accuracy, recall and F1 value. Because this method uses positional correspondence to align modern Chinese and English terms, the alignment effect in some inverted sentences is not very optimistic. In addition, the effect of ancient Chinese word segmentation will also affect the alignment accuracy of this method. In the future, we will improve this method from two aspects: the method of ancient Chinese word segmentation and the acquisition of candidate terms in inversion structure, so as to improve the effect of term alignment between ancient Chinese and English.

References
[1] Hang Z.X., (2007) English Translation of Cultural Classics and Postgraduate Teaching of Translation at Soochow University. Shanghai Translation, (1): 56-58.
[2] Wang Y.C., Tsai R.T.H., Hsu W.L. (2009) Web-based pattern learning for named entity translation in Korean–Chinese cross-language information retrieval. Expert Systems with Applications, 36(2): 3990-3995.
[3] Yang F, Zhao J., Liu K. (2009) A Chinese-English organization name translation system using heuristic web mining and asymmetric alignment. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Singapore, pp: 387-395.
[4] Zhang Y.J., Wang Y., Cen L., et al. (2010) Fusion of multiple features and ranking SVM for web-based English-Chinese OOV term translation. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING) Beijing. pp: 1435-1443.
[5] Lee C.J., Chang J.S., Shing R.J. Jang. (2006) Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources. In: Proceedings of ACM Transactions on Asian Language Information Processing (TALIP), New York, pp: 121-145.
[6] Feng D.H., Lv Y.J., Zhou M. (2004) A new approach for English-Chinese named entity alignment. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, pp: 372-379.
[7] Chen Y.F., Zong C.Q., Su K.Y. (2010) Interactive method for Chinese-English bilingual named entity recognition and alignment [J]. Journal of Computer, 34(9): 1688-1696.
[8] Chen Y.F., Zong C.P., Su K.Y. (2013) A joint model to identify and align bilingual named entities. Computational Linguistics, 39(2): 229-266.
[9] Li L.H., Wang P., Huang D.G., et al. (2011) Mining English-Chinese Named Entity Pairs from Comparable Corpora. ACM Transactions on Asian Language Information Processing (TALIP), 10(4): 1-19.
[10] You G.W., Hwang S.W., Song Y.I., et al. (2012) Efficient entity translation mining: A parallelized graph alignment approach. ACM Transactions on Information Systems (TOIS), 30(4): 1-23.
[11] Kim J.H., Hwang S.W., Jiang L., et al. (2013) Entity translation mining from comparable corpora: Combining graph mapping with corpus latent features. IEEE Transactions on Knowledge and Data Engineering, 25(8): 1787-1800.
[12] Li X.Y., Che C., Lin H.F, Wang R.P. (2010) Term translation pair alignment based on a parallel corpus of Chinese historical classics and their English translations. Journal of Information and Computational Science, 7(9): 1967-1986.
[13] Che C., Zheng X.J. (2016) Term Alignment Method in Historical Classics Based on Subwords. Journal of Chinese Information Processing, 30(3): 46-51.
[14] Wu X.T., Zhao H.Y., and Che C. (2018) Term Translation Extraction from Historical Classics Using Modern Chinese Explanation. In: Proceedings of the 17th China National Conference
on Computational Linguistics & Proceedings of the 6th International Symposium on Natural Language Processing Based on Naturally Annotated Big Data. Changsha, CCL 2018/NLP-NABD pp: 88–98.

[15] Vogel S, Ney H. (1996) HMM-based word alignment in statistical translation. In: Proceedings of the 16th International Conference on Computational Linguistics. Copenhagen, pp: 836-841.

[16] Liu Y, Liu Q, Lin S. (2005) Log-linear models for word alignment. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. University of Michigan, pp: 459-466.

[17] Wu H and Wang H.F. (2009) Revisiting Pivot Language Approach for Machine Translation. In: Proceedings of the Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the AFNLP, Singapore, pp: 154-162.

[18] Durrani N. and Koehn P. (2014) Improving Machine Translation via Triangulation and Transliteration. In: Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT). Dubrovnik, pp: 71-78.

[19] Zhu X.N., He Z.J., Wu H., Zhu C.H., Wang H.F., and Zhao T.J. (2014) Improving Pivot-Based Statistical Machine Translation by Pivoting the Co-occurrence Count of Phrase Pairs. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, pp: 1665–1676