Chinese Historical Term Translation Pairs Extraction Using Modern Chinese As a Pivot Language

Xiaoting Wu[2][0000-0002-3630-0449], Hanyu Zhao[1][0000-0002-8436-6397], Lei Jing[3], and Chao Che[1][0000-0003-2978-5430](✉)

[1] Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, Dalian University, Dalian, China
hanyuzhao7@163.com, chechao101@163.com
[2] State Grid Info & Telecom Group Beijing China-Power Information Technology Co., LTD, Beijing, China
wuxiaoting2017@163.com
[3] Chiping Vocational Education School, Liaocheng, China
jinglei.lei@163.com

Abstract. Term translation of Chinese historical classics is very difficult and time-consuming work, and using term alignment methods to extract term translation pairs is of great help for historical term translation. However, the limited bilingual corpora resources of historical classics and special morphology of the ancient Chinese result in poor performance of term alignment. To this end, this paper proposes a historical term alignment method using modern Chinese as a pivot language. The method first identifies English terms by rules, then aligns them from English to modern Chinese and then from modern Chinese to ancient Chinese. The use of English-modern Chinese corpus and modern-ancient Chinese corpus instead of English-ancient Chinese corpus solves the shortage problem of the parallel corpus. Moreover, using modern Chinese as a pivot language effectively reduces the alignment errors caused by the abbreviations and the interchangeable characters of ancient Chinese. In the term alignment experiment on ***Shiji***, our method outperformed the direct alignment method significantly, which proves the validity of our method.

Keywords: Chinese Historical Classics, Term Alignment, Pivot Language.

1 Introduction

Translating Chinese classics into English is an important way for the world to understand the history and culture of China. However, most Chinese classic books remain untranslated. At present, only about 0.2% of the 35,000 Chinese classic books have been translated[1]. One main reason for this is the translation difficulty of Chinese classic books due to the dynamic development of history and the cultural differences between China and the West. Term translation is one of the most difficult and time-consuming works in classics translation, and sometimes translators spend more than 60%

of their time on searching the proper term translation. Therefore, it will reduce the translation difficulty and save a lot of translation time to perform term alignment to extract different term translations from the bilingual corpora as the reference for the translator, which will also accelerate the translation of history books.

To the best of our knowledge, very few researches were conducted on the historical term alignment. Co-occurrence frequency[2] and maximum entropy model[3] were explored for term alignment but did not achieve satisfactory performance. Since historical terms refers to ancient official titles, numbers, institutions, apparatus, systems, events, etiquette, customs names, etc., which are similar to the named entity (NE). We investigate the NE alignment method for the term alignment research. At present, there are three main lines of bilingual NE alignment methods: (1) The symmetric method, which identifies the NEs in two languages, respectively, and then uses the alignment model to align the NEs in two sides[4]; (2) The asymmetric method, which recognizes the NEs in one language, then find its corresponding translation in another language[5-7]; (3) The integration method, which jointly perform bilingual NE alignment with other NLP tasks such as NE recognition[8,9] or word alignment[10]. However, most approaches rely on word alignment relationship to map the NEs in both languages and the performance of word alignment in historical classics are very poor due to the shortage of the parallel corpus. In addition, the special morphology of ancient Chinese further introduces many word alignment errors.

To the end, this paper proposes a pivot-based method to perform term alignment, which is an effective solution to overcome the scarceness of parallel corpus by introducing a third language that has parallel corpus with both source and target languages [11-14]. We first recognize the English terms using rule-based method, then align them from English to modern Chinese and from modern Chinese to ancient Chinese. Finally, we get the historical terms translation pairs by combing the two alignment results. We employ modern Chinese as the pivot language because most historical classic book are explained by modern Chinese and many English-modern Chinese corpora are available in public. Using modern Chinese as the pivot language can also reduce the alignment error caused by the special wording of ancient Chinese.

2 Motivation

The ancient Chinese has many lexical and syntactic differences with modern Chinese, two of which can result in many alignment errors. First, some ancient Chinese words such as people names and official names are often abbreviated while they are translated into the full name in English. Thus, it is difficult to find the corresponding English translation of the abbreviated words. For example, in the example of Figure 1, "阳成延" is abbreviated as "延" in ancient Chinese and is difficult to align with its full translation "Yang-ch'eng Yen". Second, it is very difficult to align the words containing the interchangeable characters. The interchangeable character is a character that is used to replace another character with same or similar pronunciation. Since the words containing interchangeable character usually appears in another form, it is almost impossible to find the correct translation.

Because modern Chinese and ancient Chinese share the same language system, we can address the above problems by using the corresponding modern Chinese terms. For the first problem, modern Chinese can complement the shortened terms in ancient Chinese, which make it easy to find a corresponding English translation. For example, as shown in the example in Figure 1, "延" in ancient Chinese is easily aligned with the English translation "Yang-ch'eng Yen" after mapping to the full name of "阳成延" in modern Chinese. For the second problem, modern Chinese can help identify the character that interchangeable character replaces. It is much easier to align the word containing an interchangeable character with its familiar form.

In addition, the accuracy of the word alignment between ancient and modern Chinese is very high since many words co-occur in both languages. Therefore, we can still have good alignment performance after combining two alignment results.

3 Pivot Term Alignment Method

3.1 Term Alignment Steps

In many English translations of historical books, the first letter of the words in terms are capitalized. In the same time, the ancient Chinese terms need to be identified by term recognizer, which is usually trained on tagged corpora. The English terms are much easier to identify than the ancient Chinese terms. Therefore, instead of identifying the ancient Chinese terms directly, we extract them using the term alignment. Specifically, we first identify the English term, and then perform term alignment between English and modern Chinese, the alignment between modern Chinese and ancient Chinese, respectively, and finally get the ancient Chinese terms by mapping the English terms to ancient Chinese term via modern Chinese. This procedure can be illustrated by the example in Figure 1.

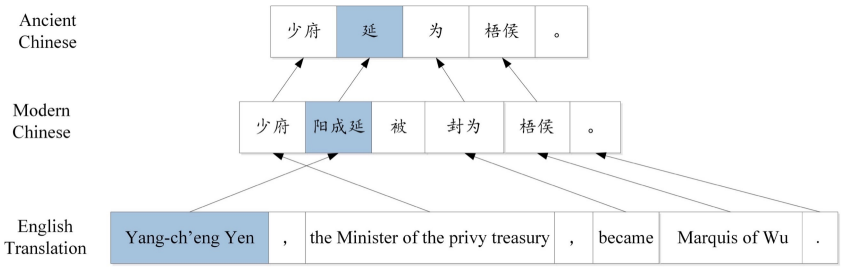

Fig. 1: An example of the term alignment using modern Chinese as the pivot language.

Given English-modern Chinese corpus $EM = \{(E_1, M_1), (E_2, M_2), \cdots, (E_N, M_N)\}$, modern-ancient Chinese corpus $MA = \{(M_1, A_1), (M_2, A_2), \cdots, (M_N, A_N)\}$, wherein E, M, A is English, modern Chinese and ancient Chinese sentence, respectively, the steps of our term alignment method can be described as Figure 2.

3.2 English Term Recognition

We make use of the capitalization rule to identify English terms since most English translations of historical books capitalize the first letter of words in the term. But the capitalization extraction rule has two problems: (1) The first word of sentence is extracted as the wrong term. (2) Some articles and conjunctions that are not capitalized in terms are missed. Thus, we make some supplementary rules for the above problems.

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, a 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.

Input: English-modern Chinese corpus EM and modern-ancient Chinese corpus MA
(1) Perform word segmentation for ancient Chinese and modern Chinese. (2) Perform word alignment between English and modern Chinese and obtain English-modern Chinese word alignment matrix A_{em}. Perform word alignment between modern Chinese and ancient Chinese and obtain modern-ancient Chinese word alignment matrix A_{ma} (3) Recognize the English terms in each English sentence E_i and get the English terms set $T_e = \{e_1, e_2, \cdots, e_n\}$ (4) Extract the corresponding modern Chinese term m_i of English term e_i according to word alignment matrix A_{em} and obtain the term translation pair set between English and modern Chinese $T_{em} = \{(e_1, m_1), (e_2, m_2), \cdots, (e_n, m_n)\}$. (5) Extract the corresponding ancient Chinese term a_i of modern Chinese term m_i according to word alignment matrix A_{ma} and obtain the term translation pair set between modern Chinese and ancient Chinese $T_{ma} = \{(m_1, a_1), (m_2, a_2), \cdots, (m_n, a_n)\}$. (6) Obtain the term translation pair set between English and ancient Chinese $T_{ea} = \{(e_1, m_1), (e_2, m_2), \cdots, (e_n, m_n)\}$ by combing the term translation pair set T_{em} and T_{ma}.
Output: English-ancient Chinese term translation pairs set T_{ea}

Fig. 2. The step of the term alignment method using modern Chinese as the pivot language.

3.3 Word Alignment

We employ IBM-4 model[15] to perform word alignment of English-modern Chinese, modern-ancient Chinese, respectively. Since the performance of word alignment model with limited scale of parallel corpora are rather unsatisfactory, there are many errors in the word alignment result. Two measures are adopted to reduce the word alignment errors:

(1) Words co-occurrence is used to improve the word alignment of modern Chinese and ancient Chinese. Specifically, if more than one character of a modern Chinese word is the same with another ancient Chinese word, they should be aligned to each other. For example, modern Chinese word "齐威王" and ancient Chinese word "齐威". If the co-occurring word pair does not appear in the word alignment matrix, they will be added. If one word of the word pair aligns to other words, we will remove the wrong word alignment and add the co-occurring word pairs.

(2) The term integrity is utilized to complement the word alignment after recognizing the English term. Specifically, all words in an English term should correspond to the same Chinese term. If some words in the English term do not map to a term, we will add the words to the alignment. For example, English term “the Lord of Hao Lake” matches Chinese term “滆池君”, while “Lord” does not align to “滆池君” in the word alignment matrix. Thus, we think the algorithm miss the alignment and add (“Lord”, “滆池君”) to the word alignment matrix.

3.4 The Bilingual Term Pairs Extraction

We conduct two kinds of term alignment to extract English-ancient Chinese term pairs, i.e., English-modern Chinese term alignment and modern Chinese-ancient Chinese alignment. Two kinds of term alignment both share the similar steps. For each term in the source sentence, we search the term in target sentence fulfilling the following conditions as the corresponding target term: (1) words of target term must be consecutive in target sentence; (2) word alignment in bilingual term pairs must be compatible with the alignment matrix, that is, the target words should align to the words in the source term or align to NULL according to the alignment matrix.

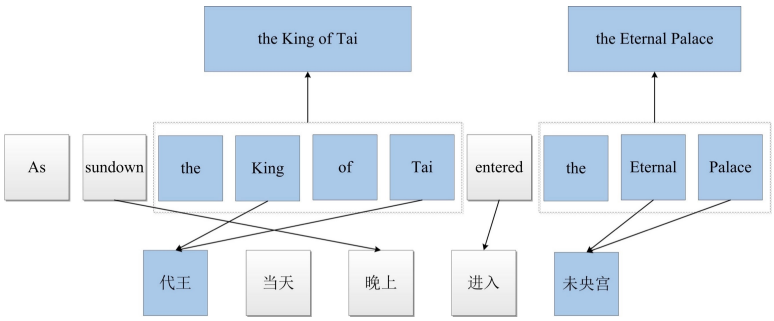

Fig. 3. A term alignment example between English and modern Chinese

We take the example shown in Fig. 3 to illustrate the term alignment process between English and modern Chinese. To match the modern Chinese term for English term "the King of Tai", we first look up the word alignment table to find the corresponding word in Chinese sentence for "King", which is "代王". Then we find the next word “of” is a preposition. We skip it and continue to match next word “Tai”, which also aligned to "代王". Therefore, words “of”, “Tai” and “the King” align to the same term "代王".

Term translation pair (“the King of Tai”, "代王") is added to the English-modern Chinese term pair set.

4 Experiment

4.1 Experimental Setup

We built three parallel corpora for the term alignment experiment, namely, English-modern Chinese corpus, modern-ancient Chinese corpus and English-ancient Chinese corpus. Each corpus comprised of 4064 sentence pairs, which are from the five basic annals of ***Shiji*** and the corresponding translation. The English translation was extracted from ***the Records of the Grand Historian of China***[16,17]. The number of term translation pairs is 1170.

In our method, word segmentation (WS) of modern Chinese was implemented by Jieba[1] and ancient Chinese was segmented by the word segmentation method based on word alignment (WSWA)[2]. The word alignment model, IBM 4 model, was implemented by GIZA++[3][18].

We employ precision(P), recall (R), and F-1 score as the evaluation metrics, which is defined as follows.

$$R = \frac{N_{\text{correct}}}{N_{gold}} \times 100\% \tag{1}$$

$$P = \frac{N_{\text{correct}}}{N_{segment}} \times 100\% \tag{2}$$

$$F1 = \frac{2PR}{P+R} \tag{3}$$

Wherein, N_{gold} is the number of words in the gold standard, N_{align} is the number of words aligned by our method; N_{corret} is the number of correctly aligned words by our method.

4.2 Experimental results and analysis

To justify the use of modern Chinese as the pivot language, we compared the proposed method with the direct alignment method, which ran IBM 4 model to align directly from English to ancient Chinese. Table 1 shows the comparison results of different term alignment methods.

[1] https://github.com/fxsjy/jieba
[2] https://github.com/supercar101/Word-Segmentation-Method-of-Ancient-Chinese/tree/master
[3] https://codeload.github.com/moses-smt/giza-pp/zip/master

Table 1. The comparison results of different term alignment methods.

Term Alignment Method	P	R	F-1
Direct alignment method	66.6%	57.8%	61.8%
Pivot based method	**81.2%**	**69.3%**	**74.8%**

In Table 1, our method shows an obvious advantage over the direct alignment method in precision, recall, and F-1 score. This contributes to the following two reasons:

(1) Using modern Chinese as the pivot language can reduce alignment errors caused by the abbreviation of ancient Chinese. The abbreviation can easily find the corresponding English translation by mapping to the full word in modern Chinese. This can be clearly illustrated by the example in Table 2.

Table 2. An example of aligning abbreviation in ancient Chinese.

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

In the ancient Chinese of table 2, terms "齐威", "楚宣", "魏惠", "燕悼" and "韩哀" all omit the titles, the full official name should be the terms in modern Chinese, i.e. "齐威王", "楚宣王", "魏惠王", "燕悼王" and "韩哀侯". When aligning ancient Chinese and English term directly, all the terms can not find the correct translation due the omission of ancient Chinese. "齐威", "楚宣" did not match English term and "魏惠", "燕悼" and "韩哀" align incorrectly to " Duke Dao ", " Duke Ai " and " Duke Cheng". In the pivot alignment, the term alignment of ancient Chinese and modern Chinese was first carried out. After finding the full name of ancient Chinese term, the term alignment between modern Chinese and English was used to find the correct translation of English terms.

(2) Modern Chinese can also help align the ancient Chinese words containing interchangeable characters to the correct translation. Interchangeable character called “通假字” in Chinese refers to a special use of ancient Chinese characters. The interchangeable characters are used to replace the characters with the same or similar pronunciation. In the direct alignment, those words containing interchangeable characters barely align to the right translation due to very low frequency. For example, ancient Chinese term "甯昌" did not find the correct translation in the direct alignment because "甯" is an interchange-

able character of "宁" and is seldom used. By aligning "甯昌" to "宁昌" in modern Chinese, we know they refer to the same name and find correct translation "Ning Ch'ang" via "宁昌".

4.3 The influence of Ancient Chinese WS

To test the influence of the ancient Chinese WS method on the term alignment, we perform term alignment using ancient Chinese WS results obtained by WSWA and the ground truth, respectively. The comparison results are shown in Table 3.

Table 3. The term alignment results using different WS results.

Methods		P	R	F-1
WSWA	Ancient Chinese WS	89.3%	83.6%	86.3%
	Term Alignment	81.2%	69.3**%**	74.8%
Ground Truth	Ancient Chinese WS	100%	100%	100%
	Term Alignment	**80.2%**	**78.8%**	**79.5%**

From Table 3, we can see Chinese WS results have a very direct influence on term alignment. The term alignment is subject to the performance limit of ancient Chinese. The recall of term alignment using the WS result is also very low. Using the ground truth can increase 9 point of recall since English terms cannot find the correct ancient Chinese term when the ancient Chinese is not segmented correctly.

Besides ancient Chinese WS, modern Chinese WS also have significant impact on the term alignment results. If the modern Chinese is segmented wrongly, English term aligns to wrong modern Chinese term and it is very hard to find the right ancient term. In the example shown in Table 4, due to the wrongly segmented modern Chinese word “缪公对”, “Duke Mu” can not find the corresponding ancient Chinese term “缪公”.

Table 4. An example of wrong modern Chinese word segmentation.

Ancient Chinese	缪公/之/怨/此/三人/入於/骨髓
Modern Chinese	缪公对/这/三个人/恨之入骨
English	The hatred Duke Mu bears these men eats into his very bones and marrow!

5 Conclusion

In this paper, we proposed a term alignment method using modern Chinese as a pivot. The method aligned the historical term from English to modern Chinese, then to ancient Chinese. Using modern Chinese as the pivot language not only solves the shortage problem of parallel corpus but also reduces the alignment error caused by abbreviation and the interchangeable characters of the ancient Chinese.

Our method only explores word alignment to perform term alignment. Currently, the performance of word alignment is far from satisfactory. This limits the performance improvement of the term alignment. In the future, we will investigate the method using more information to extract term translation pairs.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. 61402068).

References

1. Zhongxi Huang.: English Translation of Cultural Classics and Postgraduate Teaching of Translation in Suzhou University. Shanghai Journal of Translators 1, 56-58 (2007) (In Chinese).
2. Xiuying Li, Chao Che, Xiaoxia Liu, Hongfei Lin, Rongpei Wang.: Corpus-based extraction of Chinese historical term translation equivalents. International Journal of Asian Language Processing 20(2), 63-74 (2010).
3. Chao Che, Xiaojun Zheng.: The Extraction of Term Translation Pairs for Chinese Historical Classics Based on Sub-words. Journal of Chinese Information Processing 30(3), 46-51 (2016) (In Chinese).
4. Gae-Won Yout, Seung-Won Hwangt, Young-In Song, et al.: Mining name translations from entity graph mapping. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing(EMNLP), pp. 430-439. ACL, Stroudsburg (2010).
5. Donghui Feng, Yajuan Lv, Ming Zhou.: A new approach for English-Chinese named entity alignment. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 372-379. ACL, Stroudsburg (2004).
6. Chun-J en Lee, Jason S Chang, Jyh-Shing R. Jang.: Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources. ACM Transactions on Asian Language Information Processing (TALIP) 5(2), 121-145 (2006).
7. Yuejie Zhang, Yang Wang, Lei Cen, et al.: Fusion of multiple features and ranking SVM for web-based English-Chinese OOV term translation. In: 23rd International Conference on Computational Linguistics(COLING), pp. 1435-1443 ACM, New York (2010).
8. Yufeng Chen, Chengqing Zong, Keh-Yih Su.: On Jointly Recognizing and Aligning Bilingual Named Entities. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 631–639. ACL, Stroudsburg (2010).
9. Yufeng Chen, Chengqing Zong, Keh-Yih Su.: A joint model to identify and align bilingual named entities. Computational Linguistics 39(2), 229-266 (2013).
10. Mengqiu Wang, Wanxiang Che, Christopher D. Manning.: Joint Word Alignment and Bilingual Named Entity Recognition Using Dual De-composition. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 1073–1082. ACL, Stroudsburg (2013).

11. Hua Wu and Haifeng Wang.: Pivot language approach for phrase-based statistical machine translation. Machine Translation 21(3), 165-181 (2007).
12. Hua Wu and Haifeng Wang.: 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, pp. 154-162. ACL, Stroudsburg (2009).
13. Nadir Durrani and Philipp. Koehn.: Improving Machine Translation via Triangulation and Transliteration. In: Conference of the European Association for Machine Translation, pp.71-78, Dubrovnik, Croatia (2014).
14. Xiaoning Zhu, Zhongjun He, Hua Wu, Conghui Zhu, Haifeng Wang, and Tiejun Zhao.: 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), pp. 1665–1675. ACL, Stroudsburg (2014).
15. Peter E Brown, Stephen A. Della Pietra, Vincent J. Della Pietra., and Robert L. Mercer.: The mathematics of statistical machine translation: parameter estimation. Computational Linguistics, 19(2), 263-311 (1993).
16. Burton Watson.: Records of the Grand Historian of China. Columbia University Press, New York (1961).
17. Burton Watson.: Records of the Grand Historian: Qin Dynasty. Chinese University of Hong Kong, Columbia University Press, Hong Kong & New York: (1993).
18. Franz Josef Och and Hermann Ney.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29(1), 19-51 (2003).