Sentence Alignment of Historical Classics based on Mode Prediction and Term Translation Pairs

Chao Che **Xiaojun Zheng**
Key Laboratory of Advanced Design and Intelligent Computing
(Dalian University) , *Ministry of Education,*
Dalian, 116622, P. R. China
chechao101@163.com, zhengxiaojun@gmail.com

Abstract

Parallel corpora are essential resources for the construction of bilingual term dictionary of historical classics. To obtain large-scale parallel corpora, this paper proposes a sentence alignment method based on mode prediction and term translation pairs. On one hand, the method rebuilds the sentence alignment process according to characteristics of the translation of historical classics, and adds mode prediction into the sentence alignment. On the other hand, due to the lack of bilingual ancient Chinese dictionary, the method exploits the term translation pairs extracted from manually aligned sentence pairs to perform alignment. The method first predicts the alignment mode probability according to the character number, punctuation number and some characters of Chinese sentence, then performs sentence alignment using length alignment probability, term alignment probability and mode probability. Besides, the method selects anchor sentence pairs based on sentence length and predicted mode to prevent the spread of alignment errors. The experiment on "*Shi Ji*" demonstrates that mode prediction and term translation pair both enhance the performance of sentence alignment obviously.

1 Introduction

Translating the classics into English and introducing them to the world is an important way to spread Chinese culture. Because of the dynamic nature of historical development and huge differences between Chinese and Western languages, the translation of classics is very difficult. Bilingual term dictionary is very helpful for translation of historical classics. The term dictionary can be built by extracting term translation pairs from bilingual parallel corpora. Aiming at obtaining large-scale parallel corpora, we study the sentence alignment of historical classics. Currently the bilingual sentence alignment methods are mainly divided into following four types: (1) the method based on length(Gale and Church, 1993; Lu et al., 2003), which performs sentence alignment using the sentence length relation; (2) the method based on dictionary(Yu et al., 2010), which performs sentence alignment using the lexicon translation in bilingual dictionary; (3)the hybrid method(Moore, 2002; Chen and Lin, 2009; Tian et al., 2009), which makes use of many kinds of information including sentence length to improve the accuracy of alignment; (4) the method base on mode classification(Fattah et al., 2007); which regards sentence alignment as a mode classification problem and exploits classifier to align sentences.

Due to the various meanings of historical classics, the ratio of the sentence length between historical classics and its English translation is not uniform. Thus alignment method using sentence length alone does not have good performance. Owing to lack of ancient Chinese bilingual dictionary, we exploit the term translation pairs extracted from the sentence pairs aligned manually to do sentence alignment. Since the translator needs to explain the hiding meaning of historical classics when translating classics into English, the sentence alignment modes of classics are almost all "one-to-many". The sentence alignment of historical classics can be considered as finding corresponding English translation for every Chinese sentence. According to the above characteristics of historical classics, this paper proposed a sentence alignment method based on mode pre-

Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 85–89,
Wuhan, China, 20-21 October 2014

diction and term translation pairs. The method first predicts the alignment mode probability based on the features of Chinese sentence, then run sentence alignment using sentence length and term translation pairs.

2 The Mathematical Model of Sentence Alignment Method

2.1 Sentence Alignment Probability

Given a Chinese text block $C=\{c_1,c_2,\cdots,c_m\}$ and the English translation $E=e_1,e_2,\cdots,e_n$, where c_i denotes a Chinese sentence and e_j is a English sentence. Sentence alignment aims at finding the alignment $\hat{A}$ with the maximum alignment probability among all alignments *A*, which can be denoted as following formula.

$$\hat{A}=\arg\max\left\{\prod_{M\in A}\Pr(<c,e>|M(c,e))*\Pr(M(c,e))\right\} \quad (1)$$

Wherein: $<c,e>$ is a pair of sentences, which is also called sentence bead. And *e* is the translation of *c*. $M(c,e)$ indicates the alignment mode. According to the sentence number in $<c,e>$, the alignment mode can be classified as: (1:0), (0:1), (1:1), (1:2), (2:1), (1:3), (3:1) etc. Due to the rich meaning of historical classics, every sentence of classics usually corresponds to more than one English sentence. On the contrary, there is hardly any English sentence corresponds to more than one Chinese sentence. Only 4 sentence pairs are aligned in "many-to-one" mode in the corpus containing 1233 pair of sentences. Since the alignment modes of most historical classics sentences are "one-to-many", sentence alignment can be regarded as finding the corresponding English sentences $e=e_je_{j+1}\cdots$ for each Chinese sentence c_i in $C=\{c_1,c_2,\cdots,c_m\}$. Given a Chinese sentence, formula (1) is turned into:

$$\hat{A}=\arg\max_{M\in A}\left\{\Pr(e|M(c,e),c)*\Pr(M(c,e)|c)\right\} \quad (2)$$

Wherein: $\Pr(M(c,e)|c)$ is the mode probability, which denotes the probability that the alignment mode is $M(c,e)$ given the sentence c. $\Pr(e|M(c,e),c)$ indicates the probability that sentence *c* align *e* given $M(c,e)$ and *c*. $\Pr(e|M(c,e),c)$ is the linear combination of length alignment probability and term alignment probability, and is defined as following:

$$\begin{aligned}\Pr(e|M(c,e),c)=&\lambda_1\Pr_{len}(e|M(c,e),c)\\&+\lambda_2\Pr_{term}(e|M(c,e),c)\end{aligned} \quad (3)$$

Wherein: $\Pr_{len}(e|M(c,e),c)$ is the length alignment probability and $\Pr_{term}(e|M(c,e),c)$ denotes the term alignment probability. λ_1, λ_2 are the weight parameters and $\lambda_1+\lambda_2=1$.

2.2 The Length Alignment Probability

Since *c* is known, we can have the following formula not strictly.

$$\begin{aligned}&\Pr_{len}(e|M(c,e),c)\\&=\Pr_{len}(<c,e>|M(c,e))\\&=\Pr(\delta(L_c,L_e)|M(c,e))\end{aligned} \quad (4)$$

Wherein: L_c,L_e is the length of sentence *c* and *e*, respectively. Owing to the lack of mature ancient Chinese word segmentation algorithm, we take the character number as the length of Chinese sentence and take the word number as the length of English sentence. The punctuation in Chinese and English sentence is all taken into account. $\delta(L_c,L_e)$ indicates the length evaluation function, which obeys standard normal distribution, and is defined as follows.

$$\delta(L_c,L_e)=\frac{L_e-L_c\bullet c_p}{\sqrt{L_c\bullet s^2}} \quad (5)$$

Wherein: parameter c_p indicates the number of English words generated by one Chinese character in average, which can be obtained by calculating the length ratio between the paragraph P_e of English sentence *e* and the paragraph P_c of Chinese sentence *c*. The calculation formula is shown as formula (6). s^2 is the normalized factor guaranteeing $\delta(L_c,L_e)$ obey standard normal distribution, which can be obtained by calculating variance on bilingual corpus. It can be calculated as formula (7).

$$c_p=\frac{\sum_{e\in P_e}L_e}{\sum_{c\in P_c}L_c} \quad (6)$$

$$s^2=D\left(\frac{L_e-L_c\bullet c_p}{\sqrt{L_c}}\right) \quad (7)$$

Based on the 3σ principle of normal distribution, 99% values of $\delta(L_c,L_e)$ distribute in the range of [-3, 3]. To ensure the length probability

is less than 1 and incremental, length probability is defined as following:

$$\begin{aligned}&\Pr_{len}(e \mid M(c,e),c)\\&=\Pr(\delta(L_c,L_e) \mid M(c,e))\\&=1-\left|\frac{\delta(L_c,L_e)}{3}\right|=1-\left|\frac{L_e-L_c \bullet c_p}{3\sqrt{L_c \bullet s^2}}\right|\end{aligned} \quad (8)$$

2.3 Term Alignment Probability

The "terms" in historical classics refer to the official title, posthumous, person name, location name and other titles. The terms appear frequently in historical classics. If a term occurs in a Chinese sentence, then its translation should occurs in the corresponding position of English sentence. Given that a pair of terms (t_c,t_e) occur in the sentence pair (c,e), the alignment function is defined as:

$$A(t_c,t_e)=\min_{i,j=1,2,\cdots}\left|\frac{pos_i(t_c)}{L_c}-\frac{pos_j(t_e)}{L_e}\right| \quad (9)$$

Wherein: $pos_i(t_c)$ and $pos_j(t_e)$ denote the positions where terms t_c and t_e appear in the sentence. If no term pair appears in the sentence pair, the term alignment probability is 0. If *n* pair of terms $(t_{c1},t_{e1}),\cdots,(t_{cn},t_{en})$ occur in the sentence pair, the term alignment probability can defined as:

$$\begin{aligned}&\Pr_{term}(e \mid M(c,e),c)\\&=1-A_{\min}(t_{cj},t_{ej})\prod_{i\neq j}(A(t_{ci},t_{ei})+0.5)\end{aligned} \quad (10)$$

Wherein: $A_{\min}(t_{cj},t_{ej})$ is the minimum value of all the alignment functions. (t_{cj},t_{ej}) indicates the term pair when the alignment function is minimal.

2.4 Alignment Mode Probability

The historical classics are translated from Chinese to English. When translator sees a Chinese sentence, he almost determines how many English sentences should be used to translate the Chines sentence. Thus, the calculation of alignment mode probability can be considered as a classification problem of mode. We input a Chinese sentence to the classifier, and the classifier outputs the alignment probability. We take naïve Bayesian as the classifier, which is run by WEKA[1]. We employ the character number, punctuation number of Chinese sentence and the characters selected by feature selection function to predict the mode probability. We choose Information Gain (IG) as the feature function, which measures the number of bits of information obtained for category prediction by knowing the presence or absence of a character in a sentence.

3 The Framework of the Sentence Alignment Method

3.1 The steps of the sentence alignment method

Given a Chinese document D_c of historical classics and its English translation D_e, the sentence alignment is performed in the following steps.

(1) Divide the Chinese document into different paragraphs by paragraph mark, $D_c=\{C_1,C_2,\cdots,C_I\}$. The English document is also divided into different paragraphs, $D_e=\{E_1,E_2,\cdots,E_I\}$.

(2) Search anchor sentence pairs in in the paragraph pair (C_i,E_i). If anchors can be found, separate the Chinese and English paragraph into different text blocks, otherwise, regard the paragraph as a text block.

(3) Align the sentences in text blocks $c_1,c_2,\cdots,c_m$ and $e_1,e_2,\cdots,e_n$. For Chinese sentence c_i, calculate the probabilities of different alignment modes by formula (3). Select the English sentences $e_j\cdots e_{j+k}$ of the maximal alignment probability as its translation. Then find corresponding sentences for Chinese sentence c_{i+1} from English sentence e_{j+k+1}. Repeat the alignment until all the sentences in the text block are aligned.

(4) Align all the text blocks in paragraph pair (C_i,E_i). If $i+1<=I$, $i=i+1$, goto step (2), otherwise, the sentence alignment is finished.

3.2 The selection of anchor sentence pairs

We make use of anchor sentence pairs to prevent the spread of alignment errors, which often occur in the alignment method utilizing the length information. To select anchor sentence pair, we first determines Chinese anchor sentence, then

[1] http://www.cs.waikato.ac.nz/ml/weka

find English anchor sentence according to Chinese anchor sentence. To ensure the anchor sentences is correct, Chinese anchor sentence must satisfy the following four conditions: (1) The anchor sentence is not the first sentence or the last sentence of the paragraph; (2) The alignment modes of the two sentences around the anchor sentence should be predicted as (1:1); (3) The length of anchor sentence should be less than the length threshold Th_l; (4) Two sentences around anchor sentence should have Th_i more characters than anchor sentence. Condition (1) is defined because the anchor sentence has no function in alignment when it is first or last sentence. Conditions (2), (3) and (4) guarantee that corresponding sentence of anchor Chinese sentence has the smallest length in surrounding sentences. After the Chinese anchor sentence is determined, the English sentence of highest alignment probability is selected as English anchor. To enhance the computation efficiency, we do not search English anchor sentences in all sentences of the English paragraph. Instead, we find the anchor sentence in a window whose size is *wind* and whose center is at position pos_e. Position pos_e is the corresponding position of Chinese anchor sentence, calculated as formula (12).

$$pos_e = pos_c \frac{count(E)}{count(C)} \tag{11}$$

$$window = \begin{cases} 3, \text{if count(E)<=12} \\ 5, \text{if count(E)>12} \end{cases} \tag{12}$$

Wherein: $\text{count}(*)$ denotes the sentence count of the paragraph *.

4 Experiment

4.1 Experimental setup

The bilingual corpora used in the experiment are composed of *Shi Ji* and its corresponding English translations drawn from *Records of the Grand Historian*, which is well-recognized authoritative translation by famous sinologist Burton Watson. We extract 1233 sentence pairs as test corpora from 7 hereditary houses, which are *The Hereditary House of King Yuan of Ch'u, The Hereditary Houses of Ching and Yen, The Hereditary House of King Tao-hui of Ch'I, The Hereditary House of Prime Minister Hsiao,* etc. We aligned manually 4144 sentence pairs to construct the parallel corpora as training corpora from five consecutive basic annals, which are *Basic Annals of Qin*, *Basic Annals of the First Emperor of the Qin*, *Basic Annals of Hsiang Yü, Basic Annals of Emperor Kao-tsu* and *Basic Annals of Empress Lü.* We extract 641 term translation pairs from the training corpora to calculate term alignment probability. The parameters of the proposed method are set as following: weigh parameters $\lambda_1 = 0.55$, $\lambda_2 = 0.45$, length threshold $Th_l = 12$, interval threshold $Th_i = 5$.

Since we find corresponding translation for Chinese sentence one by one, almost all the sentences can be aligned, the precision and recall of our method is nearly the same. We only employ precision *p* to test the method, which is as follows.

$$P = \frac{N_{correct}}{N_{align}} \times 100\% \tag{13}$$

Wherein: $N_{correct}$ denotes the number of correct sentence pairs acquired by the proposed method, N_{align} is the number of all the sentences acquired by the proposed method.

4.2 Experimental results and analysis

The method only using sentence length is employed as baseline method. To test the effect of mode prediction and anchors, we compare the performance the method not using mode prediction, the method not using anchors with the proposed method. The precision comparison of four methods is shown in table 1.

Method	Precision
Baseline	60.5%
Not using anchors	72.2%
Not using mode prediction	86.8%
The proposed method	92.5%

Table 1: The precision comparison of four methods

Table 1 shows that both the method not using mode prediction and the method not using anchor sentence pair, which all employ term alignment probability, outperform significantly baseline method. This confirms the effectiveness of term alignment probability. It can be also seen from table 1 that the use of the anchors significantly increases the precision about 20%. The experimental results demonstrate that the anchors can effectively prevent the error spread of the alignment method based on sentence length. The result also confirms the anchor sentence pairs we obtained are correct.

In table 1, we can see that the mode prediction increases precision by 6%. In the conventional method, all the aligned sentences whether long or

short have the same mode probability. It is unreasonable since long sentences prefer “one-to-many” mode and short sentences tend to be “one-to-one” mode. The proposed method extracts the mode probability for the sentence with different features based on the training corpora, and employ different alignment probability according to features of the aligned sentence, so the precision of the proposed method is higher. However, since the model probability prediction is not very accurate, sometimes wrong mode probability leads to alignment errors. This is why the role of model prediction is not as significant as we expect.

5 Conclusion

To construct bilingual term dictionary of historical classics, this paper proposes a sentence alignment method based on mode prediction and term translation pairs. The method first obtains the mode alignment probability according to the features of Chinese sentence, then performs sentence alignment using length probability and term alignment probability. Furthermore, the method find anchor sentence pairs to prevent the spread of alignment errors. The sentence alignment experiment on “*Shi Ji*” confirms the effectiveness of the proposed method. In the future, we can further improve classification accuracy of model predictions and apply the sentence alignment in the term translation extraction of historical classics.

Acknowledgments

This work is funded the National Science Foundation of China (61402068, 61304206).

Reference

Chen, Xiang and Hong-fei Lin. 2009. Sentence Alignment of Bilingual Biomedical Abstract Based on Anchor Information. *Journal Of Chinese Information Processing*, 23(1): 58-62.

Fattah, Mohamed Abdel, David B. Bracewell, et al. 2007. Sentence alignment using P-NNT and GMM. *Computer Speech and Language*, 21: 594–608.

Gale, William A. and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1): 75-102.

Lu, Xue-qiang, Qing-yin Li, et al. 2003. Sub-Sentence Alignment of Chinese-English Law Literature Based on Stat istical Approach. *Journal of Northeastern University*, 24(1): 23-26.

Moore, Robert C. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In *AMTA*, Springer.

Tian, Shengwei, Turgun Ibrahim, et al. 2009. Chinese-Uighur sentence alignment based on hybrid strategy with mistake spread suppression. In *International Conference on Environmental Science and Information Application Technology* Wuhan, China, IEEE.

Yu, Xin, Jian Wu, et al. 2010. Dictionary-based Chinese-Tibetan sentence alignment. In *International Conference on Intelligent Computing and Integrated Systems (ICISS)*, Guilin, China.