Bilingual Sentence Alignment of Pre-Qin History Literature for Digital Humanities Study

Jiwen Liang[1], Dongbo Wang[2] and Jianlin Yang[1]

[1] Nanjing University, China dg1914006@smail.nju.edu.cn
[2] Nanjing Agricultural University, China db.wang@njau.edu.cn

Abstract. Sentence aligned bilingual text of history literature provides support of digital resources for related digital humanities studies, but existing studies have done little work on sentence alignment of ancient Chinese and English. In this study, we made a preliminary attempt to align the sentence of ancient Chinese and English. We used the bilingual text of the ***Analects of Confucius*** and ***Zuo's Commentaries of the Spring and Autumn Annals,*** extracted features and adopted the classification method to divide the bilingual candidate sentence pairs based on probability scores. The bilingual sentence alignment model based on SVM had the best performance on a larger amount of data when using three features and confirmed the impact of candidate dataset.

Keywords: Sentence Alignment, Cross-language Information Processing, Ancient Chinese and English Parallel Corpus, History Literature.

1 Introduction

With the development of computer technology and the integration of disciplines, digital humanities came into being. The digitization of history literature transforms text into electronic data forms which provides a digital platform for the combination of technology and humanities [1]. Translating of history literature and digitizing them is an important way of historical and cultural exchanges around the world. Sentence aligned bilingual text provides underlying digital resources for related digital humanities study [2]. It also has high value in the study of cross-language information retrieval and cross-language stylistic analysis. However, the following problems still exist:

- Some bilingual documents have not yet been digitized, only paper copies exist.
- Parallel units of bilingual parallel corpora for some ancient books are usually limited to text or paragraph, sentence alignment has not yet been achieved, and the information provided is limited.

This study is a part of a larger research project. In this paper, we only focused on the second question--exploring the automatic alignment of sentences in ancient Chinese and English. Over the past several decades, the study of sentence alignment has been conducted for a long period of time [3,4,5,6]. However, little work has been

done on sentence alignment of ancient Chinese and English. This study explored the following study questions:

Q1. How to automatically align ancient Chinese and English sentences?

Q2. Whether the generation pattern of candidate sentence pairs has an effect on the experiment?

2 Research Methods

In this paper, we used a combination of length and lexical information and adopted the classification method to divide the bilingual candidate sentence pairs into two categories: 'aligned sentence pairs' (labeled 'S') and 'non-aligned sentence pairs' (labeled 'O'). Assumed that the probabilities of each pair of sentences are independent, the sentence pairs with the largest probability distribution are aligned sentence pairs, the classification model was trained based on the probability scores and labels of the candidate sentence pairs.

2.1 Data Collection

After researching the existing bilingual texts of ancient Chinese-and English translation, one part of my data was the bilingual text of the ***Analects of Confucius*** which from the "Chinese Text Project " (https://ctext.org/ens) database and the other part was ***Zuo's Commentaries of the Spring and Autumn Annals***[7]which obtained by scanning paper books and OCR performed. After data cleaning and generating aligned sentences manually, we used specific punctuation as an identifier for dividing sentences. The alignment bilingual sentence pairs were combined inside the paragraph according to the different limited patterns to obtained candidate sentence pairs from two parts data as experimental dataset and assigned labels and as shown in Table 1.

Table 1. Sample of candidate sentence pair

Ancient Chinese	English	Pattern	Labels
書曰：“鄭伯克段于鄢。”	In the Annals it says：“The duke defeated Shuduan at the place of Yan.”	1-1	S
書曰：“鄭伯克段于鄢。”	Shuduan did not do his duty as a young brother, so the record omits any words about Shuduan was the young brother of the duke.	1-1	O
段不弟，故不言弟；	Shuduan did not do his duty as a young brother, so the record omits any words about Shuduan was the young brother of the duke.	1-1	S
生桓公而惠公薨，是以隱公立而奉之。	She gave birth to a son, and later he became Duke Huan. Shortly after Duke Hui died, Duke Yin ruled the state of Lu in place of Duke Huan, who was then an infant.	1-2	S
中，五之一；小，九之一。	the middle should not exceed one fifth of the capital and the small, one ninth.	2-1	S

2.2 Feature Extraction

First, we analyzed the characteristics of the bilingual texts of the pre-Qin literature. Second, we referred to the features and experimental conclusions of the existing bilingual sentence alignment researches. Overall, we excluded unusable features such as cognate words, co-occurring words and dictionaries, etc. Finally selected as three alignment features: bilingual sentence length(F1), pattern feature(F2) and bilingual keywords(F3).

In the 1990s, Church＆Gale achieved alignment of bilingual sentences based on length information [3, 4], then the calculation method has been approved and widely used by other scholars. After proving the relationship of sentence length between the ancient Chinese and English in bytes, we referred to Gale to calculated F1 and F2[2]. Considering that there are a large number of corresponding keywords of 'entities' between ancient Chinese and English (e.g., '周公'- 'duke of Zhou'). we extracted the bilingual list of keywords between ancient Chinese and English in my dataset, and calculate F3 based on the ratio of word frequency.

3 Experiment and Result

We experimented on the idea of 'overall classification' which determines the category based on the features and regardless of the classification result of the context. Experiment with **data 1** (***Analects of Confucius***) and **data 2** (***Zuo's Commentaries of the Spring and Autumn Annals***) using Support Vector Machine (SVM) [8] and Maximum Entropy Model (Maxent) respectively [9]. **Data 1** contains 1555 pairs of aligned sentences and **data 2** contains 2296 pairs. When we generated candidate pairs, **data1** used five patterns ('1-1', '1-2', '2-1', '2-2' and 'others') so the amount of data was larger, while **data 2** mainly focused on '1-1', '1-2' and the amount of data is much less than the former. After hyper parameter selection and optimization, experiments were performed on the dataset using cross-validation.

The experimental results are shown in Table 2. As the results shows:(1) SVM achieved the best result in recognizing the aligned sentence pairs of two different dataset. (2) The generation pattern of candidate sentence pairs has a great influence on the results. (3) The generation pattern of candidate sentences affected R-value when it more comprehensive.

Table 2. Experimental results

	Model	P	R	F
Data 1: *Analects of Confucius*	Maxent	**97.96%**	61.76%	75.76%
	SVM	83.46%	92.83%	**87.90%**
Data 2: *Zuo's Commentaries*	Maxent	**73.60%**	78.63%	76.03%
	SVM	71.59%	82.91%	**76.83%**

4 Discussion and Future Work

We made a preliminary attempt to align the sentence of ancient Chinese and English. The bilingual sentence alignment model based on SVM had the best performance on a larger amount of data when using three features. The analysis of results is as follows: (1) The amount of data, the generation pattern of candidate sentence pairs, and the richness of features has an impact on the results of sentence alignment. (2) The results of this study are slightly lower than previous studies such as [5-6], may be due to the large differences in the bilingual languages we used (ancient Chinese-English) with others (English-French, English-Arabic or Japanese-Chinese).

It should be noted that this study has examined only a small part of the bilingual corpus, so further exploration is needed. However, this study achieved sentence alignment in ancient Chinese and English, laying the foundation for subsequent digital humanities research such as cross-lingual retrieval and cross-lingual mining. In the future, we will increase the data from different literature and explore the best pattern for generating candidate sentence pairs. In addition, we will eventually construct a parallel corpus in the sentence-level alignment of ancient Chinese and English in Pre-Qin history literature for digital humanities study.

5 Reference

1. Holm, P., Jarrick, A., Scott, D.:The Digital Humanities. Humanities World Report 2015, Palgrave Macmillan UK (2015).
2. Guo, M., Shen, Q., Yang, Y., Ge, H., Cer, D., Abrego, G. H., ... Kurzweil, R.: Effective parallel corpus mining using bilingual sentence embeddings. arXiv preprint arXiv:1807.11906 (2018).
3. Gale, W.A, Church, K.W.: A program for aligning sentences in bilingual corpora. Computational linguistics, 75-102 (1993).
4. Church, K. W.: Char-align: a program for aligning parallel texts at the character level. In Proceedings of the 31st annual meeting on Association for Computational Linguistics, pp. 1-8. Association for Computational Linguistics (1993).
5. Fattah, M. A.: The use of MSVM and HMM for sentence alignment. Journal of Information Processing Systems, 8(2), 301-314 (2012).
6. Che, C., Guo, W., Zhang, J.: Sentence Alignment Method Based on Maximum Entropy Model Using Anchor Sentences. In Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp.76-85 (2016).
7. Luo, Zhiye.: Zuo's Commentaries of the Spring and Autumn Annals. Springer. Southeast University Publishing House, Nanjing (2015).
8. Cortes, C., Vapnik, V.: Support vector networks. Machine Learning, 20(3), 273-297 (1995).
9. Jaynes, E., T.: On the rationale of maximum-entropy methods. Proceedings of the IEEE, 70(9), 939-952 (1982).