Chinese Ancient-Modern Sentence Alignment

Zhun Lin and Xiaojie Wang

School of Information Engineering, Beijing University of Posts and Telecommunications
Beijing, China, 100876
xjwang@bupt.edu.cn

Abstract. Bi-text alignment is useful to many Natural Language Processing tasks such as machine translation, bilingual lexicography and word sense disambiguation. Most of previous researches are on different language pairs. This paper presents a diachronic alignment of Ancient and Modern Chinese. Because of the long history of Chinese culture and Chinese writing, lots of Ancient Chinese texts are waiting to be translated into modern Chinese, especially, the comparative study of Ancient and Modern Chinese is a very important way to understand some characteristics in Modern Chinese. After describing some characteristics in Ancient-Modern Chinese bi-texts, we first investigate some statistical properties of Ancient-Modern bi-text corpus, including the correlation test of text lengths between two languages and the distribution test of length ratio data. We then pay more attention to n - m ($n > 1$ or $m > 1$) alignment modes which are prone to mismatch.

1 Introduction

Text alignment is an important task in Natural Language Processing (NLP). It can be used to support many other NLP tasks. For example, it can be utilized to construct statistical translation models [1], and to acquire translation examples for example-based machine translation [2]. It can be helpful in bilingual lexicography [3].

The approaches to text alignment can be classified into two types: statistical-based and lexical-based. The statistical-based approaches rely on non-lexical information (such as sentence length, co-occurrence frequency, etc.) to achieve an alignment task [4]. The method proposed in [5] is based on the assumption that in order for the sentences in a translation to correspond, the words in them must also correspond. Their method made use of lexical anchor points to lead an alignment at the sentence level.

It has been shown that different language pairs are in favor of different information in alignment. For example, [6] found that the sentence-length correlation between English and Chinese is not as good as between English and French. Also, there is less cognate information between Chinese-English pair than that in English-French pair, while the alignment of Chinese-Japanese pair can make use of information of Hanzi commonly appearing in both languages [7]. Currently, most methods rely on either or both of above two ideas [8]. The approaches combining both length and lexical information, such as [9], seem to represent the state of the art.

Chinese has a long history. It had been changed from Ancient Chinese to Modern Chinese gradually. A lot of Ancient Chinese texts are waiting to be translated into

modern Chinese, Ancient-Modern Chinese alignment will be helpful in both example based machine translation and statistic machine translation. Although the two languages are very different in some aspects, but many of the Modern Chinese characteristics can be found in the Ancient Chinese [10], they have close relationship. Therefore, the comparative study of Ancient-Modern Chinese may give useful cues on understanding some characteristics in Modern Chinese.

In this paper, we present our current work on Ancient-Modern Chinese Sentence alignment. Combination of both length-based information and lexicon-based information is proved to be the state of art approach to bi-text alignment [8]. So we will also combine these two kinds of information in our alignment algorithm. But for Ancient-Modern pairs, as we will see, they are some special features in alignment. In Ancient-Modern pairs, Hanzi characters, which commonly occur in both languages, is obviously a useful feature for alignment, they are used for improving sentence alignment with great significance.

The remainder of paper is organized as follows: section 2 describes some characteristics in Chinese Ancient-Modern bi-text corpora, shows what we can get in a sentence level alignment. Section 3 describes several information sources we use in our approach, and how to combine these information sources for improving alignment. Where, we investigate the length correlation between two languages and the distribution of length ratio between them. We use a deliberate measure to deal with several different kinds of alignment. We then implement several experiments and evaluate different effects brought from different information sources in section 4, and give some discussions on the interesting results. Finally we draw some conclusions.

2 Some Characteristics Ancient-Modern Chinese Bi-texts

Even though Ancient Chinese and Modern Chinese are similar on lots of aspects, there are many sentence alignment modes between them. n-m (n>1 or m>1) modes, which are to be prone to mismatch, are also common in practice.

Table 1. An example of Ancient-Modern Chinese bi-text: one Ancient Chinese sentence is aligned with two Modern sentences

Mode	Ancient Chinese	Modern Chinese
1:2	徒手遇蜥蜴，变色而却步，人之情也。	两手空空，遇到一条四脚蛇，也会吓得脸上变色而后退。 这是一般人的通常心理
2:1	前日范希文贬官后，与足下相见于安道家。 足下诋诮希文为人。	前几天范希文贬官以后，我和您在安道家会面，您极力诋毁讥笑希文的为人。
2:2	故辄布区区，伏惟幸察，不宣。 修再拜。	因此就写了区区此信，恭敬地希望您明察。 不多言了，欧阳修再拜。

Some examples about various alignment modes are given in table 1. There is 1-2, 2-1 and 2-2 alignment respectively. The statistics of our data aligned manually show these kinds of alignment are more 10%, while less than 1.5% of the alignments contain 3 or more sentences in one language.

Both Ancient Chinese and Modern Chinese are composed of Hanzi Character, for example, “手” is the same character in both ancient and modern Chinese. The alignment of Ancient-Modern Chinese pairs can make use of information of Hanzi information commonly appearing in both texts.

3 The Approach to Sentence Alignment

We use dynamic programming to find overall optimal alignment paragraph by paragraph. We combine both length-based information and Hanzi-based information to measure the cost for each possible alignment between Ancient-Modern Chinese strings.

Before we give the measure function, we first define a few notations. Let s_i denote the i th sentence in Ancient Chinese, $|s_i|$ denote the number of character it includes, s_{ij} denote the string from the i th sentence to the j th sentence in Ancient Chinese text, t_u denote the u th sentence in Modern Chinese, t_{uv} denote the string from the u th sentence to the v th sentence in Modern Chinese text. $j < i$ means that s_{ij} is an empty string, and so does t_{uv} . Let $d(i, u; j, v)$ be the function that computes the cost of aligning s_{ij} with t_{uv} . We divide $d(i, u; j, v)$ into three parts, as shown in (1), to reflect that we utilize three information sources to compute the cost.

$$d(i, u; j, v) = L(i, u; j, v) + M(i, u; j, v) - \alpha H(i, u; j, v) \quad (1)$$

Where $L(i, u; j, v)$ depends on length ratio of the paired two strings s_{ij} and t_{uv} , different length ratios cause different $L(i, u; j, v)$; $M(i, u; j, v)$ depends on the alignment mode in the pair, that is, depending on how many sentences involved in this pair at both sides, different n - m modes cause different $M(i, u; j, v)$. $H(i, u; j, v)$ is the contribution from Hanzi characters common in both strings.

We first describe how to use length-based information, and then Hanzi-based information. We give a deliberate measure on multiple alignments which are prone to mismatch.

3.1 Length and Mode

To utilize the length information in Ancient-Modern Chinese alignment, we make an investigation on length correlation between these two languages. We also check if the normal hypothesis of the length ratio is appropriate for this language pair. Following Gale and Church (1991), we use alignment data at paragraph-level to do the correlation and normal test. 205 paragraphs in our Ancient-Modern corpus are randomly chosen for these tests.

The correlation test is to examine if the lengths of Ancient Chinese texts are related to their Modern translations. That is, if longer texts in Ancient Chinese tend to be translated into longer texts in Modern Chinese. Figure 1 shows that the lengths (in characters) of Ancient and Modern Chinese paragraphs are highly correlated.

Fig. 1. Paragraph lengths are highly correlated in Ancient-Modern Chinese bi-texts. The horizontal axis shows the length of Ancient Chinese paragraphs, while vertical scale shows the lengths of the corresponding Modern Chinese paragraphs.

Fig. 2. The length-rate data of Ancient-Modern Chinese bi-texts is approximately normal. The horizontal axis shows the length-ratio, while the vertical axis shows the frequency of each ratio.

In [4], they based their distance measure on an assumption that each character in one language gives rise to a random number of characters in other language. They assume those random variables are independent and identically distributed with a normal distribution. They check the English-French pair for this assumption, and estimate parameters for the normal distribution. This assumption is not clear to Ancient-Modern Chinese pair. For making use of the same kind of length-based information in

alignment, we thus give an investigation on if this assumption can be held in Ancient-Modern Chinese pair. We also use the paragraph-based data to do the test. We use 205 length ratios from 205 Ancient-Modern Chinese paragraphs as samples. Figure 2 is the histogram for these ratios.

We can find that length ratio data is approximately normal with mean $\mu=1.6$ and Std. Deviation $\sigma=0.21$. Then following [4], we compute $L(i,u;j,v)$ using equation (2).

$$L(i,u;j,v)=-100*\log(2*(1-\Pr(|t_{uv}|/|s_{ij}|-\mu)/\sigma)) \tag{2}$$

To compute the cost for different alignment mode, we aligned 407 Ancient-Modern Chinese sentence pairs manually. Then we estimate the probabilities of different alignment modes using frequencies of these modes occurring in data. The result is listed in Table 2.

Supposed that $j-i=n$ means the string s_{ij} including n Ancient Chinese sentences, specially, we use $n=-1$ to denote that there is no sentence in s_{ij} . Similarly, when $v-u=m$ means the string t_{uv} including m Modern Chinese sentences and $m=-1$ means that there is no sentence in t_{uv} , the cost for mode $n-m$ can be computed by equation (3).

$$M(i,u;j,v)=-100*\log(\Pr(n-m)) \tag{3}$$

Table 2. the frequencies of different alignment modes

Mode(n-m)	Frequency f(n-m)	Probability Pr(n-m)
0-1, 1-0	1	0.002
1-1	350	0.861
1-2	21	0.052
2-1	18	0.044
2-2	11	0.027
Others	6	0.014
Total	407	1

3.2 Hanzi Information

Another source of information is Hanzi occurring commonly in both Ancient and Modern Chinese. We include totally 2366 Hanzi character pairs in a dictionary. Following [7], we use different measures as in equation (4) and (5) for different alignment mode. The equation (5) is for 2-2 mode, and the equation (4) is for other modes. By including some length-related information in these measures, we can also control the alignment length.

$$H(i,u;j,v)=\begin{cases} 0 & j < i \vee v > u & 0-1 \text{ or } 1-0 \\ h_1(i,0;u,0) & j = i \ \& \ v = u & 1-1 \\ h_1(i,0;u,0) + h_2(i,j;u,0) & j = i + 1 \ \& \ v = u & 2-1 \\ h_1(i,0;u,0) + h_3(i,0;u,v) & j = i \ \& \ v = u + 1 & 1-2 \end{cases} \tag{4}$$

Where $h_1(i, 0, u, 0) = |s_i \cap t_u| / \min\{|s_i|, |t_u|\}$, here we use the relative number of commonly occurring Hanzi to measure the similarity between two strings, $h_2(i, j, u, 0) = (|s_j \cap t_u| - |s_i \cap s_j \cap t_u|) / |s_j|$, which is the relative number of commonly occurring Hanzi only in s_j and t_u but not in s_i , similarly, we use $h_3(i, 0, u, v)$ to denote $(|s_i \cap t_v| - |s_i \cap t_u \cap t_v|) / |t_v|$, which is the relative number of commonly occurring Hanzi only in s_i and t_v but not in t_u . For 2-2 mode, we use (5) as a measure.

The three items in the right of equation (5) are the measures for three different ways of merging s_i and s_j and matching with t_u and t_v in Chinese-Japanese alignment.

$$H(i, u; j, v) = \min \{ \begin{aligned} &h_1(i, 0; u, 0) + h_2(i, j; u, 0) + h_1(0, j; 0, v), \\ &h_1(i, 0; u, 0) + h_3(i, 0; u, v) + h_1(0, j; 0, v), \\ &h_1(i, 0; 0, v) + h_1(0, j; u, 0) \end{aligned} \quad (5)$$

This completes the description of distance measure. We finally use dynamic programming to find the overall least cost in matching. Let $D(i, u)$ be total distance aligning sentences from first one to i th in Ancient Chinese and sentences from first one to u th in Modern Chinese, the recurrence then can be described in (6).

$$D(j, v) = \min\{D(i-1, u-1) + d(i, j; u, v)\} \quad (6)$$

4 Experiments and Evaluations

We use 205 Chinese Ancient-Modern paragraph pairs in our experiments, which include total 1528 Ancient Chinese sentences and 1583 Modern sentences respectively.

To estimate α parameter in equation 1, we align another 53 paragraphs which include 440 Ancient Chinese sentences and 452 Modern Chinese sentences. We use it as training data in our experiments when α is needed.

Our purpose of this experiment was to test a variety of information sources in the sentence alignment of Ancient-Modern Chinese. Table 3 is the results.

Table 3. How different factors effect the accuracy(+: used, -: not used)

H	L	M	Accuracy(%)
+	-	-	80.33
+	+	-	56.01
+	-	+	92.43
+	+	+	79.59
-	+	-	52.09
-	-	+	52.50
-	+	+	78.70

We can find Hanzi information(H) is the most helpful one for alignment, when only one information source is used, H achieves the best accuracy.

It is interesting that we can get the best result by combining only H with M instead of combining all three information sources in other language pairs such as Chinese-Japanese [7]. We think the reason is our H measure combines enough length constraint information, while length information itself is not so helpful in Ancient-Modern Chinese alignment.

We can find from Table 3 than length information give the worst performance. when only the length information is considered, the alignment result inclines to match the length rate with 1.6. When length rates locate in symmetrical position of a normal distribution, for example 1.4 and 1.8, they have the same probability. In this case, length information is not helpful in distinguish the situation in Figure 3, where 1-1 and 2-1 alignment have same probability.

Fig. 3. one length rate is 1.4 (1-1), the other is 1.8(2-1), they have the same probability

When only the alignment mode information is considered, the alignment result more inclined to match the 1-1 mode.

But when there are more sentences in Modern Chinese, things will be changed, for example, when there is one Ancient Chinese sentence and two Modern Chinese sentences left. Two alignment sets can be possible, one is (1:2), another s (1:1 and 0:1). the costs of these two sets are listed in Table 4. As we can see, the cost of (1-1,0-1) mode is larger than others, so 1-2 mode is used.

Table 4. Two kinds of alignment: 1-1,0-1(or 0-1, 1-1) and 1-2 mode

Mode	Probability Pr(n-m)	cost
1-1,0-1	0.861, 0.002	276
1-2	0.052	128

5 Conclusions

This paper describes an Ancient-Modern Chinese alignment at the level of sentence by combining both length information and Hanzi character information. We give a detail description on some characteristics in Ancient-Modern Chinese bi-texts. We check the correlation between the lengths of Ancient Chinese text and Modern Chinese text, and

find that length ratio data fits normal hypothesis approximately. We pay a special attention on n - m alignment where n or m is greater than 1, and use a similarity measure based on Hanzi information to get better accuracy than normal one. Experiments show our proposal is very helpful.

We will extend our program to manage clause alignment in our future works. By analyzing the sentence alignment, we find that there are less alignment modes in clause alignment than that in sentence alignment. Therefore, we can expect to achieve higher accuracy for Ancient-Modern Chinese clause alignment.

References

- [1] Peter F. Brown, Jennifer C. Lai, Robert L. Mercer: Aligning Sentences in Parallel Corpora. *Proceedings of 29th Annual Conference of the Association for Computational Linguistics* (1991) 169-176.
- [2] Hiroyuki Kaji, Yuuko Kida, Yasutugu, Morimoto: Learning Translation Templates from Bilingual Text. In *Proceedings of the fifteenth International Conference on Computational Linguistics*, Nantes, France (1992) 672-678.
- [3] Jörg Tiedemann: Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing, Doctoral Thesis. *Studia Linguistica Upsaliensia 1*, ISSN 1652-1366, ISBN 91-554-5815-7.
- [4] William A. Gale, Kenneth W. Church: A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of 29th Annual Conference of the Association for Computational Linguistics* (1991) 177-184.
- [5] Martin Kay, Martin Roscheisen: Text-Translation Alignment. *Computational Linguistics*. Vol. 19, No. 1 (1993) 121-142.
- [6] Wu Dekai: Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. In *Proceedings of the 32st Annual Meeting of the Association for Computational Linguistics*, (1994) 80-87.
- [7] Wang Xiaojie, Ren Fuji: Chinese-Japanese Clause Alignment. *Computational Linguistics and Intelligent Text Processing, 6th International Conference* (2005) 400-412
- [8] Jean Veronis: From the Rosetta stone to the information society—A survey of parallel text processing. In *Parallel Text Processing*. J. Veronis (eds.) (2000) 25-47 Kluwer Academic Publishers. Printed in Netherland.
- [9] I. Dan Melamed: Pattern recognition for mapping bitext correspondence. In *Parallel Text Processing*, J. Veronis (eds.), Kluwer Academic Publishers. Printed in Netherland. (2000)25-47
- [10] Changxu Sun: Introduce to Ancient-Chinese lexicon, *Shanghai dictionary Press* (2005)(in Chinese)