Aligning Bilingual Corpora Using Sentences Location Information*

Li Weigang Liu Ting Wang Zhen Li Sheng
Information Retrieval Lab, Computer Science & Technology School,
Harbin Institute of Technology 321#
Harbin, China, 150001
{LEE, tliu, wangzhen, lis}@ir.hit.edu.cn

Abstract

Large amounts of bilingual resource on the Internet provide us with the probability of building a large scale of bilingual corpus. The irregular characteristics of the real texts, especially without the strictly aligned paragraph boundaries, bring a challenge to alignment technology. The traditional alignment methods have some difficulties in competency for doing this. This paper describes a new method for aligning real bilingual texts using sentence pair location information. The model was motivated by the observation that the location of a sentence pair with certain length is distributed in the whole text similarly. It uses (1:1) sentence beads instead of high frequency words as the candidate anchors. The method was developed and evaluated through many different test data. The results show that it can achieve good aligned performance and be robust and language independent. It can resolve the alignment problem on real bilingual text.

1 Introduction

There have been a number of papers on aligning parallel texts at the sentence level in the last century, e.g., (Brown et al. 1991; Gale and Church, 1993; Simard et al. 1992; Wu DeKai 1994). On clean inputs, such as the Canadian Hansards and the Hong Kang Hansards, these methods have been very successful.

(Church, Kenneth W, 1993; Chen, Stanley, 1993) proposed some methods to resolve the problem in noisy bilingual texts. Cognate information between Indo-European languages pairs are used to align noisy texts. But these methods are limited when aligning the language pairs which are not in the same genre or have no cognate information. (Fung, 1994) proposed a new algorithm to resolve this problem to some extent. The algorithm uses frequency, position and recency information as features for pattern matching. (W. Bin, 2000) adapted the similar idea with (Fung, 1994) to align special domain bilingual texts. Their algorithms need some high frequency word pairs as anchor points. When processing the texts that include less high-frequency words, these methods will perform weakly and with less precision because of the scarcity of the data problem.

(Haruno and Yamazaki, 1996) tried to align short texts without enough repeated words in structurally different languages, such as English and Japanese. They applied the POS information of content words and an online dictionary to find matching word pairs. But this is only suitable for the short texts.

The real text always includes some noisy information. It has the following characteristics as follows:

1) There are no strict aligned paragraph boundaries in real bilingual text;

2) Some paragraphs may be merged into a larger paragraph because of the translator's individual idea;

3) There are many complex translation patterns in real text;

4) There exist different styles and themes;

5) Different genres have different inherent characteristics.

The tradition approaches to alignment fall into two main classes: lexical and length. All these methods have limitations when facing the real text according to the characteristics mentioned above.

* This research was supported by National Natural Science Foundation (60203020) and Science Foundation of Harbin Institute of technology (hit.2002.73).

We proposed a new alignment method based on the sentences location information. Its basic idea is that the location of a sentence pair with certain length is distributed in the whole text similarly. The local and global location information of a sentence pair is fully combined together to determine the probability with which the sentence pair is a sentence bead.

In the first of the following sections, we describe several concepts. The subsequent section reports the mathematical model of our alignment approach. Section 4 presents the process of anchors selection and algorithm implementation is shown in section 5. The experiment results and discussion are shown in section 6. In the final section, we conclude with a discussion of future work.

2 Several conceptions

1) Alignment anchors: (Brown, 1991) firstly introduced the concept of alignment anchors when he aligned Hansard corpus. He considered that the whole texts were divided into some small fragments by these alignment anchors. Anchors are some aligned sentence pairs.

2) Sentence bead: and at the same time, (Brown, 1991) called each aligned sentence pair a sentence bead. Sentence bead has some different styles, such as (0:1), (1:0), (1:1), (1:2), (1: more), (2:1), (2:2), (2: more), (more: 1), (more: 2), (more: more).

3) Sentence pair: Any two sentences in the bilingual text can construct a sentence pair.

4) Candidate anchors: Candidate anchors are those that can be possible alignment anchors. In this paper, all (1:1) sentence beads are categorized as candidate anchors.

3 Mathematical Model of Alignment

The alignment process has two steps: the first step is to integrate all the origin paragraphs into one large paragraph. This can eliminate the problem induced by the vague paragraph boundaries. The second step is the alignment process. After alignment, the bilingual text becomes sequences of translated fragments. The unit of a fragment can be one sentence, two sentences or several sentences. The traditional alignment method can be used with the fragment with several sentences to improve the alignment granularity. In this paper the formal description of the alignment task was given by extending the concepts of bipartite graph and matching in graph theory.

3.1 Bipartite graph

Bipartite graph: Here, we assumed G to be an undirected graph, then it could be defined as G=<V, E>. The vertex+ set of V has two finite subsets: V_1 and V_2, also $V_1 \cup V_2 = V$, $V_1 \cap V_2 = \phi$. Let E be a collection of pairs, when $e \in E$, then $e=\{v_i, v_j\}$, where $v_i \in V_1, v_j \in V_2$. The triple G was described as, $G=<V_1, E, V_2>$, called bipartite graph. In a bipartite graph G, if each vertex of V_1 is joined with each vertex of V_2, or vice versa, here an edge represents a sentence pair. The collection E is the set of all the edges. The triple $G=<V_1, E, V_2>$ is called complete bipartite graph. We considered that: $|V_1|=m$, $|V_2|=n$, where the parameters m and n are respectively the elements numbers of V_1 and V_2. The complete bipartite graph was usually abbreviated as Km, n (as shown in figure 1).

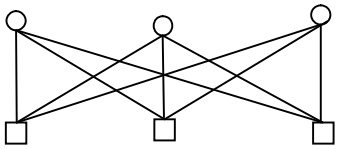

Figure 1 $K_{3,3}$ complete bipartite graph

3.2 Matching

Matching: Assuming $G=<V_1, E, V_2>$ was a bipartite graph. A matching of G was defined as M, a subset of E with the property that no two edges of M have a common vertex.

3.3 Best Alignment Matching

The procedure of alignment using sentence location information can be seen as a special matching. We defined this problem as "Best Alignment Matching" (BAM).

BAM: If $M=<S, E_M, T>$ is a best alignment matching of G=<S, E, T>, then M must meet the following conditions:

1) All the vertexes in the complete bipartite graph are ordered;

2) The weight of any edges in E_M $d(s_i, t_j)$ has: $d(s_i, t_j) < D$ (where D is alignment threshold); at the same time, there are no edges $\{s_k, t_r\}$ which made k<i and r>j, or k>i and r<j;

3) If we consider: |S|=m and |T|=n, then the edge $\{s_m, t_n\}$ belonged to E_M;

Best alignment matching can be attained by searching for the smallest weight of edge in collection E, until the weight of every edge $d(s_i, t_j)$ is equal or more than the alignment threshold D. Generally, the alignment threshold D is determined according to experience because different texts have different styles.

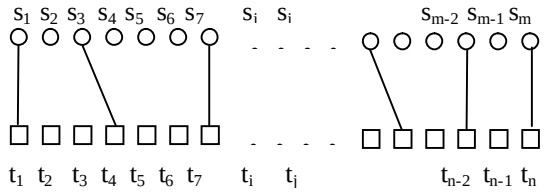

Figure 2 Sketch map of $K_{m, n}$ BAM under alignment threshold D

If each sentence in the text S (or T) corresponds with a vertex in V_1(or V_2), the text S or T can be denoted by $S(s_1, s_2, s_3,\ldots s_i, \ldots s_j, \ldots s_m)$ or $T(t_1, t_2, t_3\ldots t_i, \ldots t_j, \ldots t_n)$. Considering the form merely, each element in S combined with any element in T can create a complete bipartite graph. Thus the alignment task can be seen as the process of searching for the BAM in the complete bipartite graph. As shown in figure 2, the edge $e = \{s_i, t_j\}$ belongs to M; this means that the i-th sentence in text S and the j-th sentence in text T can make an alignment anchor. Each edge is corresponding to an alignment value. In order to ensure the bilingual texts are divided with the same fragment number, we default that the last sentence in the bilingual text is aligned. That is to say, $\{s_m, t_n\} \in E_M$ was correct, if |S|=m and |T|=n in the BAM mathematical model.

We stipulated the smaller the alignment value is, the more similar the sentence pair is to a candidate anchor. The smallest value of the sentence pair is found from the complete bipartite graph. That means the selected sentence pair is the most probable aligned (1:1) sentence bead. Alignment process is completed until the alignment anchors become saturated under alignment threshold value.

Sentence pairs extracted from all sentence pairs are seen as alignment anchors. These anchors divide the whole texts into short aligned fragments. The definition of BAM ensures that the selected sentence pairs cannot produce cross-alignment errors, and some cases of (1: more) or (more: 1) alignment fragments can be attained by the fragments pairs between two selected alignment anchors.

4 Anchors Selection during Alignment

All (1:1) sentence beads are extracted from different styles of bilingual texts. The distribution states that all of them are similar as presented in figure 3. The horizontal axis denotes the sentence number in Chinese text, and the vertical axis denotes the sentence number in English text.

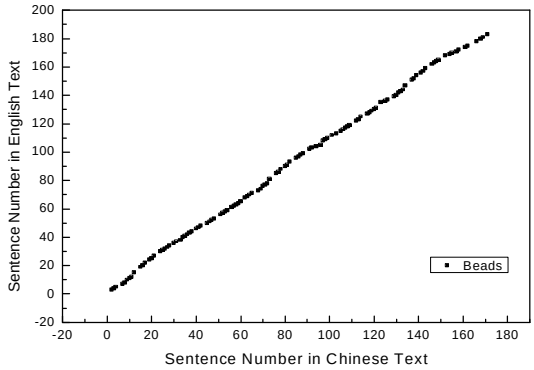

Figure 3 Distribution of (1:1) sentence beads in bilingual texts

Statistical results show that more than 85% sentence beads are (1:1) sentence beads in bilingual texts and their distributions obey an obvious law well. (DeKai Wu, 1994) offered that (1:1) sentence beads occupied 89% in English-Chinese as well. If we select these style sentence beads as candidate anchors, the alignment method will be general on any other language pairs. The main points of our alignment method using sentences location information are: locating by the whole text, collocating by sentence length and checking by a bilingual dictionary. Location information of any sentence pair is used fully. Three lengths are used: are sentence length, upper context length above the sentence pair and nether context length below the sentence. All this information is considered to calculate the alignment weight of each sentence pair. Finally, the sentence pair with high weight will be checked by a English-Chinese bilingual dictionary.

In order to study the relationship between every sentence pair of $\{s_i, t_j\}$, four parameters are defined:

Whole text length ratio: $P0 = Ls / Lt$;

Upper context length ratio: $Pu[i, j] = Us_i / Ut_j$;

Nether context length ratio: $Pd[i, j] = Ds_i / Dt_j$

Sentence length ratio: $Pl[i, j] = Ls_i / Lt_j$;

Where

s_i the i-th sentence of S;

t_j the j-th sentence of T;

Ls the length of source language text S;

Lt the length of target language text T;

Ls_i the length of s_i;

Lt_j the length of t_j;

Us_i the upper context length above sentence s_i;

Ut_j the upper context length above sentence t_j;

Ds_i the nether context length below sentence s_i;

Dt_j the nether context length below sentence t_j;

Figure 4 illustrates clearly the relationship of all variables.

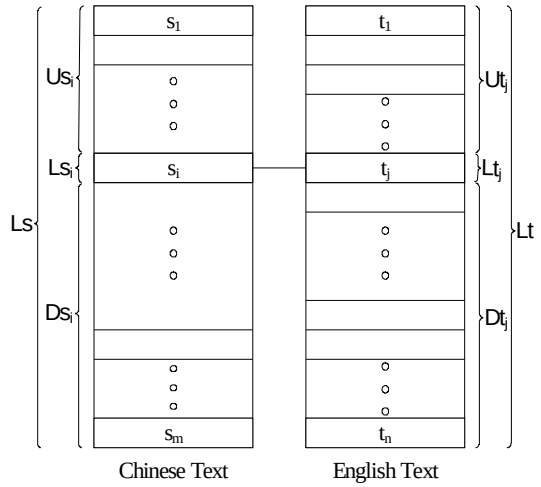

Figure 4 Sketch map of variables relationship

If s_i and t_j can construct a (1:1) alignment anchor, P[i, j] must be less than the alignment threshold, where P[i,j] denotes the integrated alignment value between s_i and t_j. We assume that the weight coefficient of Pl[i, j] is 1. Only considering the form, Pu[i, j] and Pd[i, j] must have the same weight coefficient. Here the weight coefficient is set α. We constructed a formal alignment function on every sentence pair:

P[i,j] =

$\alpha(Pu[i, j]-P0)^2 + (Pl[i, j] -P0)^2 +\alpha(Pd[i, j] -P0)^2$

Where, the parameter α is the weight coefficient, if can adjust the weight of sentence pair length and the weight of context lengths well. The longer the text is, the more insensitive the effect of the context-length is. So α's value should increase in order to balance the whole proportion. The short text is vice versa. In this paper we define:

$\alpha= (Ls/Ls_i + Lt/Lt_j)/2$

According to the definition of BAM, the smaller the alignment function value of P[i, j] is, the more the probability of sentence pair $\{s_i, t_j\}$ being a (1:1) sentence bead is. In this paper, we adopt a greedy algorithm to select alignment anchors according to all the alignment function values of P[i, j] which are less than the alignment threshold. This procedure can be implemented with a time complexity of O(m*n).

To obtain further improvement in alignment accuracy requires calculation of the similarity of the sentence pairs. An English-Chinese bilingual dictionary is adopted to calculate the semantic similarity between the two sentences in a sentence pair. The similarity formula based on a bilingual dictionary is followed:

$$H = \frac{L|Match(S)|+L|Match(T)|}{L|S|+L|T|}$$

Where L| | is the bytes number of all elements, Match (T) is (according to English-Chinese dictionary) the English words which have Chinese translation in the Chinese sentence, Match (S) is the matched Chinese fragments.

According to the above dictionary check, alignment precision is improved greatly. We take a statistic on all the errors and find that most errors are partial alignment errors. Partial alignment means that the alignment location is correct, but a half pair of the alignment pairs is not integrated. It is very difficult to avoid these errors when only taking into account the sentence location and length information. Thus in order to reduce this kind of error, we check the semantic similarity of the context-adjacent sentence pairs also. Because these pairs could be other alignment patterns, such as (1:2) or (2:1), the similarity formulas have some difference from the (1:1) sentence pair formula. Here, a simple judgement is performed. It is shown as:

If（$Ls_{i-1} * P0 > Lt_{j-1}$）

$$H_{\text{adjacent}} = \frac{L|Match(S)|+L|Match(T)|}{(1+P0)*Ls_{\text{adjacent}}}$$

else

$$H_{\text{adjacent}} = \frac{L|Match(S)|+L|Match(T)|}{(1+1/P0)*Lt_{\text{adjacent}}}$$

Here, those alignment anchors whose similarities exceed the similarity threshold based on the bilingual dictionary will become the final alignment anchors. These final anchors divide the whole bilingual texts into aligned fragments.

5 Algorithm Implementation

According to the definition of BAM, the first selected anchor will divide the whole bilingual texts into two parts. We stipulated that the sentences in the upper part of source text cannot match any sentence in the nether part of target text. As shown in Fig 5, after the first alignment anchors were selected, the second candidate anchors must be selected in the first quadrant or the third quadrant and exclusive from the boundary. It is obvious that if the candidate anchors exist in the second quadrant or fourth quadrant, the cross alignment will happen. For example, if the (i, j) is the first selected alignment anchor, and the (i-1, j+1) is the second selected alignment anchor, the cross alignment appears. We can limit the anchors selection field to prevent the cross-alignment errors.

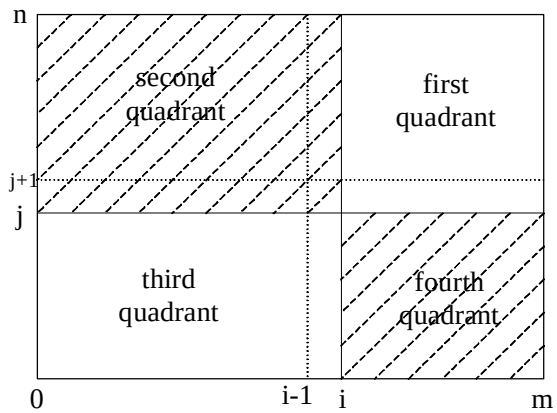

Figure 5 Anchors selection in Bilingual Texts

In addition, in order to resolve the problem that the first sentence pair is not a (1:1) sentence bead, we use a virtual sentence length as the origin alignment sentence bead when we initialize the alignment process.

The implementation of alignment algorithm is described as followed:

1) Load the bilingual text and English-Chinese dictionary;

2) Identify the English and Chinese sentences boundaries and number each sentence;

3) Default the last sentence pair to be aligned and calculate every sentence pair's alignment value;

4) Search for sentence pair that is corresponding to the smallest alignment function value;

5) If the smallest alignment function value is less than the alignment threshold and the go to step 6), and if the smallest value is equal to or more than the threshold, then go to step 7);

6) If the similarity of the sentence pair is more than a certain threshold, the sentence pair will become an alignment anchor and divide the bilingual text into two parts respectively, then limit the search field of the next candidate anchors and go to the step 4)

7) Output the aligned texts, and go to the end.

6 Results and Discussion

We use the real bilingual texts of the seventeenth chapter in the literary masterpiece "Wuthering Heights" as our test data. The basic information of the data is shown in the table 1.

English text size	38.1K
Chinese text size	25.1K
English sentence number	273
Chinese sentence number	277

Table 1 Basic information of the test data

In order to verify the validity of our algorithm, we implement the classic length-based sentence alignment method using dynamic programming. The precision is defined:

Precision = The correct aligned sentence pairs / All alignment sentence pairs in bilingual texts

The comparison results are presented in table 2.

Method	Precision (%)
Length-based alignment method	20.3
Location-based alignment method	87.8

Table 2 Comparison results between two methods

Because the origin bilingual texts have no obvious aligned paragraph boundaries, the error extension phenomena happen easily in the length-based alignment method if the paragraphs are not strictly aligned correctly. Its alignment results are so weaker that it cannot be used. If we omit all of the origin paragraphs information, we merge all the paragraphs in the bilingual text into one larger paragraph respectively. The length-based alignment method rated the precision of 25.4%. This is mainly because the English and Chinese languages don't belong to the same genre and have large difference between the language pairs. But our

method rated 129 (1:1) sentence pairs as alignment anchors which divide the bilingual text into aligned fragments. The length-based classic method was applied to these aligned fragments and got a high precision. Fig 6 shows 129 selected anchors distribution which is in the same trend with all the (1:1) sentence beads. Their only difference is the sparse extent of the aligned pairs.

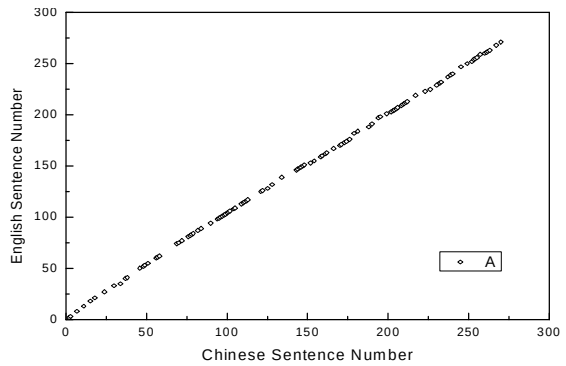

Figure 6 Distribution of alignment anchors

In order to evaluate the adaptability of our method, we select texts with different themes and styles as the test set. We merge two news bilingual texts and two novel texts. The data information is show in Table 3.

Our method is applied on the fixed data and receives the precision rating of 86.9%. The result shows that this alignment method is theme independent.

English text size	63.9K
Chinese text size	41.5K
English sentence number	510
Chinese sentence number	526

Table 3 Basic information of the fixed test data

(Haruno and Yamazaki, 1996) tried to align short texts in structurally different languages, such as English and Japanese. In this paper the aligned language pairs of English and Chinese belongs to structurally different languages as well. Our method gets the highest precision in aligning short texts. A bilingual news text is selected to be test data. The result is shown in table 4. There are two aligned sentence error pairs which are induced by the lack the corresponding translation.

English text size	5.6K
Chinese text size	3.4K
English sentence number	40
Chinese sentence number	38
Precision (%)	94.4

Table 4 Alignment results of short test data

It is difficult to attain large test set because doing so need more manual work. We construct the test set by merging the aligned sentence pairs in the existing sentence aligned bilingual corpus into two files. Then the two translated files can be as test set. Here we merge 2000 aligned sentence pairs. The file information is as follows:

English text size	200.3K
Chinese text size	144.2K
English sentence number	2069
Chinese sentence number	2033

Table 5 Basic information of the large test data

From the table 4, it is evident that there are many different styles of sentence beads. The method is developed on this large test set and gets the precision of 90.5%. The reason of the slight precision increase is that the last test set is relatively clean and the sentence length distribution relatively average. But overall, our method performs very well to align the real bilingual texts. It shows the high robustness and is not related to the languages, text themes, text length. This method can resolve the alignment problem of the real text.

7 Conclusion

This paper proposed a new method for fully aligning real bilingual texts using sentence location information, described concretely in section 3 and 4. The model was motivated by the observation that the location of a sentence pair with certain length is distributed in the whole text similarly. It uses the (1:1) sentence beads instead of the high frequency words as the candidate anchors. Local and global location characteristics of sentence pairs are involved to determine the probability which the sentence pair is an alignment anchors.

Every sentence pair corresponds to an alignment value which is calculated according to the formal alignment function. Then the process of BAM is performed to get the alignment anchors. This alignment method can restrain the errors extension effectively in comparison to the traditional alignment method. Furthermore, it has shown strong robustness, even if when it meets ill-quality texts

that include incorrect sentences. To obtain further improvement in alignment accuracy sentence similarity based on an English-Chinese dictionary was performed. It need not segment the Chinese sentence. The whole procedure requires little cost to implement.

Additionally, we can adjust the alignment and similarity thresholds dynamically to get high precision alignment anchors, for example, applying the first test set, even if we get only 105 (1:1) sentence beads but the precision is 100%. We found that this method can perform the function of paragraph alignment very well and ensure simultaneous the alignment precision.

Of these pairs about half of total number of (1:1) sentence beads can be even extracted from the bilingual text directly to build a large scale bilingual corpus if the original bilingual text is abundant. And the rest bilingual text can be used as spare resource. Now, we have obtained about 500,000 English-Chinese aligned sentence pairs with high quality.

In the future, we hope to do further alignment on the basis of current work and extend the method to align other language pairs.

References

Wu, DeKai. 1994. *Aligning a parallel English-Chinese corpus statistically with lexical criteria*. In Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics, 80--87, Las Cruces, New Mexico

Simard, M., Foster, G., and Isabelle, P. 1992. *Using Cognates to Align Sentences in Bilingual Corpora*.Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92), Montreal, Canada

Brown, P., Lai, J. and Mercer, R. 1991. *Aligning Sentences in Parallel Corpora*. ACL-91

Fung Pascale and Kathleen Mckeown. 1994. *Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping*. In AMTA-94, Association for Machine Translation in the Americas, 81--88, Columbia, Maryland

Wang Bin, Liu Qin, Zhang Xiang. 2000. Automatic Chinese-English Paragraph Segmentation and Alignment. *Journal of Software*, 11(11):1547-1553 (Chinese)

Church, Kenneth W. 1993. *Char_align: A Program for Aligning Parallel Texts at the Character Level.* Proceedings of ACL-93, Columbus OH

Chen, Stanley. 1993. *Aligning Sentences in Bilingual Corpora Using Lexical Information*. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-1993)

Gale, W.A. Church, K.W. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(2): 75-102

Haruno, Masahiko & Takefumi Yamazaki (1996), *High-performance bilingual text alignment using statistical and dictionary information*, In *Proceedings of ACL '96*, Santa Cruz, California, USA, pp. 131-138

M. Kay & M. Roscheisen. 1993. Text-Translation Alignment. *Computational Linguistics* 19:1