Combining Sentence Length with Location Information to Align Monolingual Parallel Texts

Weigang Li, Ting Liu, and Sheng Li

Iformation Retrieval Laboratory, School of Computer Science and Technology, Box 321, Harbin Institute of Technology, Harbin, P.R. China, 150001,
{lee,tliu,lis}@ir.hit.edu.cn,
WWW home page: http://ir.hit.edu.cn/

Abstract. Abundant Chinese paraphrasing resource on Internet can be attained from different Chinese translations of one foreign masterpiece. Paraphrases corpus is the corpus that includes sentence pairs to convey the same information. The irregular characteristics of the real monolingual parallel texts, especially without the strictly aligned paragraph boundaries between two translations, bring a challenge to alignment technology. The traditional alignment methods on bilingual texts have some difficulties in competency for doing this. A new method for aligning real monolingual parallel texts using sentence pair's length and location information is described in this paper. The model was motivated by the observation that the location of a sentence pair with certain length is distributed in the whole text similarly. And presently, a paraphrases corpus with about fifty thousand sentence pairs is constructed.

1 Introduction

Paraphrases are alternative ways to convey the same information [1]. And the paraphrase phenomenon is a common language phenomenon. The paraphrasing technology has been applied for various applications of natural language processing, such as question answering [2–6], information extraction [7], machine translation [8], information retrieval [9], multidocument [10], and it can improve the whole performance of these applications. There are many kinds of naturally occurred paraphrases resource. Barzilay extracted them from different translations of the same masterpiece [1]. Shinyama extracted paraphrases from news articles [11]. But the language they processed most is English or Japanese, the Chinese paraphrases is seldom researched. And in this paper, we mainly referred to the Barzilay's method to build a Chinese paraphrases corpus. The practical alignment technology is necessary for doing it.

Alignment technology is mainly applied for the bilingual texts. There have been a number of papers on aligning bilingual texts at the sentence level in the last century, e.g., [12–15]. On clean inputs, such as the Canadian Hansards and the Hong Kang Hansards, these methods have been very successful. Church [16] and chen [17] 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 languages pairs which are not in the same genre or have no cognate information. Fung [18] proposed a new algorithm to resolve this problem to some extent. The algorithm uses frequency, position and recency information as features for pattern matching. Wang [19] adapted the similar idea with Fung [18] to align special domain bilingual texts. Their algorithms need some high frequency word pairs as features. 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.

The real monolingual parallel texts always include some noisy information. They have the following characteristics as follows:

1) The monolingual parallel texts are aligned in chapter;

2) There are no strict aligned paragraph boundaries in real monolingual parallel text;

3) Some paragraphs may be merged into a larger paragraph since the translator's individual idea;

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

5) There exist different styles and themes.

The tradition approaches to alignment fall into two main classes: lexical and length. All these methods have limitations when facing the real monolingual parallel texts according to the characteristics mentioned above. We proposed a new alignment method based on the sentences length and location information. The 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, 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

It is necessary to clarify several concepts for understanding the alignment process. As shown below:

1) Alignment anchors: Brown [12] firstly introduced the concept of alignment anchors when he aligned Hansard corpus. He considered that anchors are some aligned sentence pairs which divided the whole texts into small fragments.

2) Sentence bead: And at the same time, Brown [12] called each correct 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 monolingual parallel 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 Alignment Mathematical Model

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 monolingual parallel texts become sequences of aligned fragments. And the unit of a fragment can be one sentence, two sentences or several sentences.

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 $V_1 \cup V_2 = V$, $V_1 \cap V_2 = \emptyset$. Let E be a collection of pairs, when $e \in E$, then e=$\{v_i, v_j\}$, where $vi \in V_1$,$vj \in V_2$. 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.

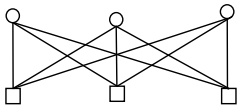

Fig. 1. $K_{3,3}$ complete bipartite graph

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 length and location information can be seen as a special matching. We defined this problem as "Best Alignment Matching" (BAM).

BAM: If $M =< S, EM, T >$ is a best alignment matching of $G =< S, E, T >$, then E_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(si, tj) has: d(si, tj)< D (where D is alignment threshold); at the same time, there are no edges sk, tr 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 sm, tn 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(si, tj) is equal or more than the alignment threshold D. Generally, the alignment threshold D is determined according to experience.

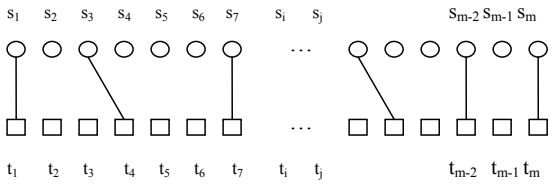

Fig. 2. Sketch map of $K_{m,n}$ BAM under alignment threshold D

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(s1, s2, s3, ..., si, sj, ..., sm) or T(t1, t2, t3, ..., ti, tj, ..., tn). 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 = {si, tj} belongs to E_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 monolingual parallel texts are divided with the same fragment number, we default that the last sentence in the monolingual parallel texts is aligned. That is to say, {sm, tn}$\in E_M$ 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 be 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. At

the same time, these anchors themselves are extracted as correct sentence pairs independently. 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 Alignment Anchors Selection

All (1:1) sentence beads are extracted from different styles of monolingual parallel texts. Their distribution states are similar as presented in Figure 3. The horizontal axis denotes the sentence number in one Chinese translation text, and the vertical axis denotes the sentence number in another Chinese translation text.

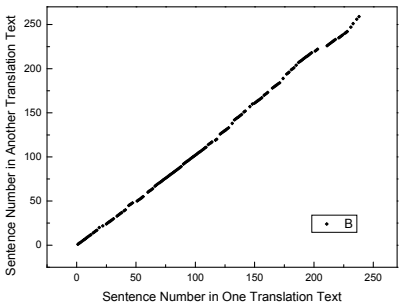

Fig. 3. Distribution of (1:1) sentence beads in Monolingual Parallel texts

Statistical results show that more than 85% sentence beads are (1:1) sentence beads in monolingual parallel texts and their distributions obey an obvious law well. DeKai, Wu offered that (1:1) sentence beads occupied 89% in English-Chinese as well [15]. If we select these sentence beads as candidate anchors, the alignment method will be general on any languages pairs. Length and location information of sentence pair is used fully to calculate the alignment weight of each sentence pair. Finally, the sentence pair with high value will be filtered by the similarity of the two sentences in a sentence pair.

In order to calculate the alignment value of sentence pair of si, tj, four parameters are defined:

Whole text length ratio: P0 = Ls / Lt;

Upper context length ratio: Pu[i, j] = Usi / Utj;

Nether context length ratio: Pd[i, j] = Dsi / Dtj

Sentence length ratio: Pl[i, j] = Lsi / Ltj;

Where

si the i-th sentence of S;

tj the j-th sentence of T;

Ls the length of one translation text S;

Lt the length of anther translation text T;
Lsi the length of si;
Ltj the length of tj;
Usi the upper context length above sentence si;
Utj the upper context length above sentence tj;
Dsi the nether context length below sentence si;
Dtj the nether context length below sentence tj;
Figure 4 illustrates clearly the relationship of all variables.

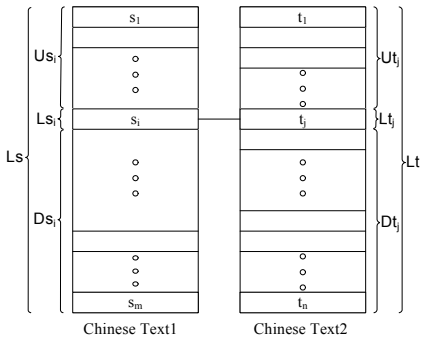

Fig. 4. Sketch map of variables relationship

If si and tj 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 si and tj. We construct a formal alignment function on every sentence pair:

$$P[i,j] = \alpha(Pu[i,j] - P0)^2 + (Pl[i,j] - P0)^2 + \alpha(Pd[i,j] - P0)^2 + (1/Sim)$$

Where, the parameter α is balanced coefficient, and it can adjust the weight of sentence pairs' 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 change in order to balance the whole proportion. The short text is vice versa. In this paper we define:

$$\alpha = (Ls/Lsi + Lt/Ltj)/2$$

And the similarity of the two sentences in a sentence pair is calculated through:

$$Sim = \frac{Intersection(L_1, L_2) \times 2}{L_1, L_2} \times \frac{min(L_1, L_2)}{max(L_1, L_2)}$$

Where L_1 and L_2 are the lengths of the two sentences; min(L_1,L_2) is the length of the shorter sentence, max(L_1,L_2) is the length of longer sentence; Intersection(L_1,L_2) is the common length of the two sentences.

According to the definition of BAM, the smaller the alignment function value of P[i, j] is, the more the probability of sentence pair si, tj 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 the similarity is used to filter the wrong sentence pairs independently. And calculation approach of the similarity is same with the method mentioned above.

After the above similarity filtering, although the alignment recall is reduced, the alignment precision is improved greatly. Here, those candidate alignment anchors whose similarities exceed the similarity threshold will become the final alignment anchors. These final anchors divide the whole monolingual parallel texts into aligned fragments.

5 Algorithm Implementation

According to the definition of BAM, the first selected anchor will divide the whole monolingual parallel texts into two parts. We stipulated that the sentences in the upper part of one translation text cannot match any sentence in the nether part of anther translation text. As shown in Figure 5.

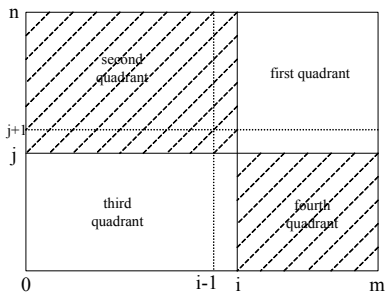

Fig. 5. Anchors selection in Monolingual Parallel texts

After the first alignment anchor was selected, the second candidate anchor must be selected in the first quadrant or the third quadrant and exclusive from the boundary. It is obvious that the cross alignment error will happen if the candidate anchor exists in the second quadrant or fourth quadrant. 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.

In addition, in order to resolve the problem that the first and the last 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 monolingual parallel texts;

2) Identify the sentences boundaries and number each sentence;

3) Calculate every sentence pair's alignment value;

4) Search the 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 monolingual parallel 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

Because the translations of most of masterpiece are aligned in chapter and the sentence number of every chapter are less than 500, our algorithm works well on the monolingual parallel texts with the sentence number under 500. Part of translations in "The Sorrows of Young Werther" and "Cien Años de Soledad" are selected as test set. The concrete information is shown in Table 1 and 2.

Table 1. Concrete information of the test data

		Translation 1	Translation 2
The Sorrows of Young Werther	Size	18.40K	18.40K
	Sentence number	238	260
Cien Años de Soledad	Size	49.5K	49.2K
	Sentence number	639	559

Table 2. Distribution of different alignment style

	Total sentence number	1:1	1:2	2:1	other
The Sorrows of Young Werther	222	185	24	8	5
Cien Años de Soledad	520	436	44	19	21

The alignment experiments are performed under the condition of with similarity filtering and without similarity filtering. The precision and recall are defined:

Precision = The correct aligned sentence pairs number / The total number of alignment sentence pairs in monolingual parallel texts

Recall = The correct aligned sentence pairs number / The total alignment sentence pairs in standard test texts

The comparison results are presented in Table 3.

Table 3. Comparison results with similarity

		Precision(%)	Recall(%)
The Sorrows	Without similarity filtering	70.9	75.1
of Young Werther	With similarity filtering	85.5	72.3
Cien Aïos de	Without similarity filtering	68.5	74.2
Soledad	With similarity filtering	83.2	70.6

With similarity filtering, the 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 pair is not integrated. The result shows that similarity filtering can resolve the problem in some extent. The recall is so low because there are some correct aligned fragment pairs with more than one sentence which cannot match the correspondent sentence pair in the standard set.

In order to verify the validity of our algorithm, we implement the classic length-based sentence alignment method using dynamic programming. And combining the traditional alignment method with our method, the results are shown in Table 4.

Table 4. Comparison results between two methods

	Precision(%)	Recall(%)
Method based on length	35.0	33.9
Our method	85.5	72.3
Combination method	91.2	85.6

Because the origin monolingual parallel texts have no obvious aligned paragraph boundaries, the error extension phenomena happen easily in the length-based alignment method. Its alignment results are so weaker that it cannot be used. If we omit all of the origin paragraphs information and merge all the paragraphs in the monolingual parallel text into one larger paragraph respectively. The length-based alignment method rated the precision of 35.0%. This is mainly because different translators have different translation styles and different comprehension on the same foreign texts. But our method rated 160 (1:1) sentence pairs as alignment anchors which divide the monolingual parallel text into aligned fragments. Then the length-based classic method was applied to these aligned fragments and got a high precision.

Figure 6 shows 160 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. We can make a conclusion that Our method performs very well to align the real monolingual parallel texts.

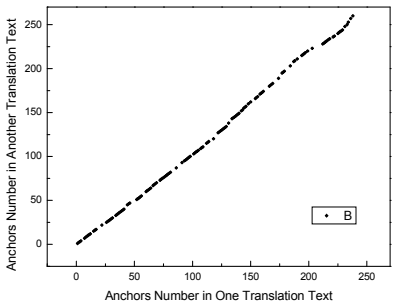

Fig. 6. Distribution of alignment anchors

7 Conclusion

This paper proposed a new method for fully aligning real monolingual parallel texts using sentence length and 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 length-based 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 filtering was performed. The algorithm need not segment the Chinese sentence require 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 107 (1:1) sentence beads but the precision is 98.13%. We found that this method can perform the function of paragraph alignment very well and ensure the alignment precision simultaneously.

Of these pairs about half of total number of (1:1) sentence beads can be even extracted from the monolingual parallel text directly to build a large scale paraphrase corpus if the original monolingual parallel text is abundant. And the rest text can be used as spare resource. Now, we have obtained about 50,000 Chinese paraphrase 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 languages pairs.

8 Acknowledgments

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

References

1. Barzilay, R., McKeown, K.: Extracting paraphrases from a parallel corpus. In: Meeting of the Association for Computational Linguistics. (2001) 50–57
2. Lin, D., Pantel, P.: Discovery of inference rules for question answering. Natural Language Engineering **1** (2001)
3. Rinaldi, F., Dowdall, J., Kaljurand, K., Hess, M., Mollá, D.: Exploiting paraphrases in a question answering system. In Inui, K., Hermjakob, U., eds.: Proceedings of the Second International Workshop on Paraphrasing. (2003) 25–32
4. France, F.D.: Learning paraphrases to improve a question-answering system. In: EACL-Natural Language Processing for Question Answering. (2003)
5. Tomuro, N.: Interrogative reformulation patterns and acquisition of question paraphrases. In Inui, K., Hermjakob, U., eds.: Proceedings of the Second International Workshop on Paraphrasing. (2003) 33–40
6. Takahashi, T., Nawata, K., Kouda, S., Inui, K., Matsumoto, Y.: Effects of structural matching and paraphrasing in question answering. IEICE Transactions on Information and Syste (2003)
7. Shinyama, Y., Sekine, S.: Paraphrase acquisition for information extraction. In Inui, K., Hermjakob, U., eds.: Proceedings of the Second International Workshop on Paraphrasing. (2003) 65–71
8. Kanayama, H.: Paraphrasing rules for automatic evaluation of translation into Japanese. In Inui, K., Hermjakob, U., eds.: Proceedings of the Second International Workshop on Paraphrasing. (2003) 88–93
9. Jacquemin, C.: Syntagmatic and paradigmatic representations of term variation. In: 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), Proceedings, Maryland, pages (1999) 341–348
10. Barzilay, R., Elhadad, N., McKeown, K.R.: Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research **17** (2002) 35–55
11. Shinyama, Y., Sekine, S., Sudo, K., Grishman, R.: Automatic paraphrase acquisition from news articles (2002)
12. Brown, P.F., Lai, J.C., Mercer, R.L.: Aligning sentences in parallel corpora. In: Meeting of the Association for Computational Linguistics. (1991) 169–176
13. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. Computational Linguistics **19** (1993) 75–102
14. Simard, M., Foster, G.F., Isabelle, P.: Using cognates to align sentences in bilingual corpora. In: Proc. of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT, Montreal, Canada (1992) 67–81
15. Wu, D.: Aligning a parallel english-chinese corpus statistically with lexical criteria. In: Meeting of the Association for Computational Linguistics. (1994) 80–87
16. Church, K.W.: Char_align: A program for aligning parallel texts at the character level. In: ACL93. (1993) 1–8

17. Chen, S.F.: Aligning sentences in bilingual corpora using lexical information. In: Meeting of the Association for Computational Linguistics. (1993) 9–16
18. PASCALE, F., MCKEOWN, K.: Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping (1994)
19. Bin, W., Qin, L., Xiang, Z.: Automatic chinese-english paragraph segmentation and alignment. Journal of Software **11** (2000) 1547–1553