Proceedings of the 15th COLING,
1994, pp. 166-171.

A Part-of-Speech-Based Alignment Algorithm

Kuang-hua Chen and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.
e-mail: hh_chen@csie.ntu.edu.tw

Abstract

To align bilingual texts becomes a crucial issue recently. Rather than using length-based or translation-based criterion, a part-of-speech-based criterion is proposed. We postulate that source texts and target texts should share the same concepts, ideas, entities, and events. Simulated annealing approach is used to implement this alignment algorithm. The preliminary experiments show good performance. Most importantly, the experimental objects are Chinese-English texts, which are selected from different language families.

1. Introduction

Real texts provide the alive phenomena, usages, and tendency of language in a particular space and time. This recommends us to do the researches on the corpora. Recently, many researchers further claim that "two languages are more informative than one" (Dagan, 1991). They show that two languages could disambiguate each other (Gale *et al*., 1992); bilingual corpus could form a bilingual dictionary (Brown *et al*., 1988) and terminology correspondence bank (Eijk, 1993); a refined bilingual corpus could be formed the examples for machine translation systems (Sumita *et al*., 1990). To do such kinds of researches, the most important task is to align the bilingual texts.

Many length-based alignment algorithms have been proposed (Brown *et al*., 1991; Gale and Church, 1991a). The correct rates are good. However, the languages they processed belong to occidental family. When these algorithms are applied to other running texts from different families, will the performance keep on the same level? Other translation-based alignments (Kay, 1991; Chen, 1993) show the difficulty in determining the word correspondence and are very complex.

In this paper, we will introduce a part-of-speech (POS)-based alignment algorithm. Section 2 will touch on the level of alignment and define the sentence terminators. In Section 3, we will propose the criterion of critical POSes and investigate the distribution of these POSes in the Chinese-English texts. Section 4 will describe a fair and rigorous method for evaluating performance. Then, we apply simulated annealing technique to conducting experiments and show the experimental results in Section 5. Section 6 will give a brief conclusion.

2. Alignment Problem

Alignment has three levels: 1) paragraph; 2) sentence; and 3) word. Paragraph level is sometimes called discourse level. Many efforts are involved in sentence level and fewer researchers touch on the word level (Gale and Church, 1991b). To do sentence alignment, we should first define what a sentence is. In English, the sentence terminators are full stop, question mark, and exclamation mark. However, the usage of punctuation marks is unrestricted in Chinese and the types of punctuation marks are numerous (Yang, 1981). Nevertheless, in order to parallel the languages, we define that the sentence markers are full stop, question mark, and exclamation mark over all languages. Therefore, an alignment of two texts is to find a best sequence of sentence groups, which are ended with one of the sentence terminators.

Following Brown *et al.* (1991), we use the term *bead*. A bead contains some sentences of source and target texts. Thus, alignment can be defined as (1).

(1) An alignment is to find a bead sequence under some criteria.

If the applied criteria are significant, the performance will be good. Finding significant criteria is the core of this research.

3. Criteria of Alignment

Any alignment algorithm has its own criteria. For example, many alignment algorithms are based on sentence length and word correspondence. Here, we propose a POSes-based criterion.

(2) Alignment Criterion:
The numbers of critical part of speeches (POSes) of a language pair in an aligned bead are close.

Now, the problem is what forms the critical POSes. Following many grammar formalisms (Sells, 1985), the content words will be the good indicators. Therefore, we think nouns, verbs, and adjectives as the critical POSes. In addition, we include numbers and

quotation marks in the critical POSes due to intuition. The English tagging system used in this work follows that of the LOB Corpus (Johansson, 1986). The Chinese tagging system follows that of the BDC Corpus (BDC, 1992) but with some modifications. The BDC Corpus does not assign tags to punctuation marks. We adopt the same philosophy of LOB Corpus to assign the tags of the punctuation marks as themselves. These critical POSes in English and in Chinese are listed in Table 1. N- represents all tags initiated with N, i.e., - is a wildcard.

Table 1. Critical POSes

	English (LOB tags)	Chinese (BDC tags)
Noun	N-, P-	n-
Verb	V-, H-, B-, D-	v
Adjective	A-	a
Number	CD-, OD-	q
Quotation Marks	*', **'	「, 」, 『, 』, 《, 》

Our bilingual corpus is investigated to check the effectiveness of the postulation (2). Ten aligned Chinese-to-English texts, CE_01 to CE_10, are considered as the objects of experiments. These texts are selected from Sinorama Magazine, published in Chinese and English monthly by Government Information Office of R.O.C. Appendix lists the source of these ten texts. We compute the average of differences (AD), variance of differences (VD), and standard deviation of differences (SD) of the critical POSes. Table 2 itemizes the values.

Table 2. Statistics of Bilingual Texts

Files		CE_01	CE_02	CE_03	CE_04	CE_05	CE_06	CE_07	CE_08	CE_09	CE_10	Total
Bead Leng.		28	39	29	33	27	23	25	53	61	37	355
Noun	AD	1.857	2.103	2.069	2.758	2.192	2.043	2.360	2.509	2.759	2.541	2.379
	VD	2.694	4.348	3.375	5.093	7.078	2.824	5.990	5.986	5.218	3.438	4.811
	SD	1.641	2.085	1.837	2.257	2.661	1.681	2.448	2.447	2.284	1.854	2.193
Verb	AD	1.000	1.153	1.724	2.333	1.615	1.522	1.680	2.189	1.776	1.757	1.758
	VD	1.429	4.045	2.614	3.434	1.852	1.728	1.898	2.266	2.277	1.968	2.531
	SD	1.195	2.011	1.617	1.853	1.361	1.314	1.378	1.505	1.509	1.403	1.591
Adj.	AD	0.929	1.462	2.310	1.061	1.308	2.391	0.880	1.679	2.379	1.459	1.644
	VD	0.781	1.941	2.697	0.845	1.444	3.282	0.586	2.067	3.684	2.032	2.372
	SD	0.884	1.393	1.642	0.919	1.202	1.812	0.765	1.438	1.919	1.426	1.540
Num.	AD	0.321	0.949	0.655	0.636	0.769	1.217	0.600	0.736	0.500	0.838	0.707
	VD	0.290	1.690	0.502	0.777	0.716	1.474	0.720	0.949	0.733	1.055	0.948
	SD	0.538	1.300	0.708	0.881	0.846	1.214	0.849	0.974	0.856	1.027	0.974
Quot.	AD	0.357	0.282	0.345	0.303	0.923	0.609	0.520	0.481	0.207	0.432	0.416
	VD	0.587	0.356	0.778	0.514	1.917	1.195	2.010	1.193	0.509	0.840	0.949
	SD	0.766	0.597	0.882	0.717	1.385	1.093	1.408	1.092	0.713	0.917	0.974
Total	AD	2.607	3.487	3.724	2.788	3.654	2.826	3.640	4.358	3.690	3.297	3.496
	VD	3.667	11.173	12.614	4.955	8.534	7.709	10.790	14.607	10.559	5.182	9.714
	SD	1.915	3.343	3.552	2.226	2.921	2.776	3.285	3.820	3.249	2.276	3.117

For oriental languages like Chinese, the correspondence with its aligned counterpart in occidental languages is not so manifest like alignment of two languages within the same family. On the one

hand, the POSes may be changed within an alignment; on the other hand, a sequence of English words may correspond to a Chinese word. These phenomena make the alignment much harder. However, Table 2 shows that these POSes are good indicators for alignment.

4. How to Evaluate the Performance

Before proceeding the experiment, another important issue is how to evaluate the correct rate for an alignment. No literatures touch on this issue. We just find in literatures what performance is rather than how to evaluate it. Note that alignment has the order constraint. On the one hand, when an error occurs, the performance should drop quickly. On the other hand, the error will not broadcast to the next paragraph. That is to say, the error will be limited in a range. Our criterion for evaluating performance takes care of the two factors.

For a given text, we could manually find the real alignment. This alignment consists of a sequence of beads, as mentioned previously. We call the sequence of beads *Real Bead Sequence (RBS)*. In contrast, we may apply any alignment algorithm to finding an alignment. We call this alignment *Computed Bead Sequence (CBS)*.

In order to evaluate the performance of the alignment algorithm, we further define the *Incremental Bead Sequence (IBS)*.

(3) *Incremental Bead Sequence IBS* of a given bead sequence *BS* is a bead sequence, such that bead IB_i in *IBS* is summation of B_j $(0 \le j \le i-1)$ in *BS*.

Therefore, two possible *IBS*es, *IRBS* and *ICBS*, are generated under this consideration. We define performance for an alignment as

(4) Performance = $\dfrac{\text{number of common beads in IRBS and ICBS}}{\text{number of beads in IRBS}}$

Table 3 demonstrates how to calculate the performance. Two beads, (3,3) and (10,10), are shared by *IRBS* and *ICBS*. The total number of beads in *IRBS* is 10. Therefore, the performance is 20%. In the following experiment, we will use this method to evaluate performance.

Table 3. Examples for Calculating Performance

RBS	(1,0), (1,1), (1,2), (1,1), (1, 2), (1,1), (1,1), (1,0), (1,1), (1,1)
IRBS	(1,0), (2,1), **(3,3)**, (4,4), (5,6), (6,7), (7,8), (8,8), (9,9),**(10,10)**
CBS	(1,1), (1,1), (1,1), (1,0), (1,1), (1,1), (1,2), (2,1), (1,2)
ICBS	(1,1), (2,2), **(3,3)**, (4,3), (5,4), (6,5), (7,7), (9,8), **(10,10)**
performance	2/10 = 0.2 = 20%

5. Alignment Algorithm

The alignment algorithms proposed in the past literatures try to find an optimal alignment which has the largest alignment probability. Due to the very large search space, they all consider only five types of beads: (0,1), (1,1), (1,2), (2,1), and (1,0). After examining our corpus, we can find other types of beads such as (1,3) and (1,4). Furthermore, bead type (2,4) is also found. Table 4 lists the distribution of bead types in the testing texts. Eight bead types appear in the bilingual texts. Bead type (1,1) is the majority (63.9%). Bead types (1,3), (1,4), and (2,4), which are not treated in other papers, occupy 8.2%. If the alignment algorithm did not deal with these bead types, the correct rate would be bound to 91.8%.

Table 4. Distribution of Bead Types

Bead Type	(1,0)	(0,1)	(1,1)	(2,1)	(1,2)	(1,3)	(1,4)	(2,4)	Total
#	2	2	227	6	89	20	8	1	355
%	0.56	0.56	63.94	1.69	25.07	5.63	2.25	0.28	100

It shows the difficulty of the alignment task. If we allow various types of beads and adopt the optimal search, the processing cost is too high to stand. A

good algorithm should satisfy the following two conditions:

- It is a general local search algorithm.
- It allows the unlimited bead types in the aligning process.

Under this consideration, simulated annealing approach (Aarts and Korst, 1989) is used to align texts. The idea of annealing comes from condensed matter physics. It involves two steps: 1) increasing temperature of matter; 2) decreasing temperature gradually until the matter in the ground configuration. Simulated annealing is to simulate the annealing process. Therefore, a simulated annealing mechanism is composed of four parts: configuration, transition function, energy function, and annealing schedule. If we take an alignment as a configuration, the possible alignments constitute the configuration space. In addition, every configuration is associated with an energy. The optimal configuration is the one which has the lowest energy. Simulated annealing is to find the optimal configuration from an initial configuration by generating a sequence of configurations under a control parameter.

For our application, we introduce another component, Transition Vector. The five components are defined as follows.

(5) *Configuration* (C): An alignment is a configuration naturally. For example, a possible bead sequence, {(1,2), (1,1), (1,1), (1,2), (1,1)}, is a configuration.

(6) *Transition Function* (T): Given a configuration, this function is responsible for generating its next configuration. A transition vector is generated at random, and then the transition function moves one configuration to another configuration according to the transition vector.

(7) *Transition Vector* (TV): A transition vector consists of 4 components (*B, N, W, D*).
B denotes the identification (counted from 0) of a selected bead.

N specifies whether to generate a new bead or not. If *N* equals to 0, no new bead is generated. If *N* equals to 1, a new bead is generated.

W represents which language in the selected bead should be moved out. If *W* equals to 0, one of the marginal sentences of the first language should be moved out. Otherwise, one of the marginal sentences of the second language should be moved out.

D represents the moving direction. 0 denotes the left marginal sentence of the selected bead is moved left, and 1 denotes the right marginal sentence of the selected bead is moved right.

For example, transition function will transit a configuration {(1,2), (1,1), (1,1), (1,2), (1,1)} to {(1,2), (1,1), (0,1), (1,0), (1,2), (1,1)} according to the transition vector TV = (2, 1, 0, 1).

(8) *Energy Function* (E): Assume each sentence has a weight, which is measured by the number of critical POSes. The weight difference of a bead is the difference between the weights of respective sentences in one bead. The energy of a configuration is the sum of weight differences of all beads in a configuration.

(9) *Annealing Schedule* (AS): When a new configuration *C'* is generated, two alternatives are considered: move to the new configuration *C'* or retain the current configuration *C*. The criterion is if *E(C')* < *E(C)*, the new configuration is adopted. However, if

$$\exp\left(\frac{E(C) - E(C')}{cp_k}\right) > random\,[0,1)$$

we will also move to the new configuration. Otherwise, the current configuration is retained. This is the well-known *Metropolis Criterion*. The cp_k is the control parameter, which will be reduced gradually in the annealing process.

Now, we apply the simulated annealing to aligning the texts, CE_01 to CE_10. The initial control parameter cp_k is 1.0 and initial run length L_k is 1000. We reduce the control parameter with 0.5% after each run. The initial configuration is randomly generated. We conduct two experiments, 1) without using paragraph markers; 2) with using paragraph markers. The results are shown in Table 5 and Table 6, respectively.

Table 5. Correct Rate for Simulated Annealing (without using paragraph marker)

Texts	CE_01	CE_02	CE_03	CE_04	CE_05	CE_06	CE_07	CE_08	CE_09	CE_10	Total
Correct	25	28	23	26	21	20	22	39	49	27	280
Total	28	39	29	33	27	23	25	53	61	37	355
Correct Rate	0.893	0.718	0.793	0.788	0.778	0.870	0.880	0.736	0.803	0.730	0.789

Table 6. Correct Rate for Simulated Annealing (with using paragraph marker)

Texts	CE_01	CE_02	CE_03	CE_04	CE_05	CE_06	CE_07	CE_08	CE_09	CE_10	Total
Correct	28	36	28	30	26	23	25	49	55	35	335
Total	28	39	29	33	27	23	25	53	61	37	355
Correct Rate	1.000	0.923	0.966	0.909	0.963	1.000	1.000	0.925	0.902	0.946	0.944

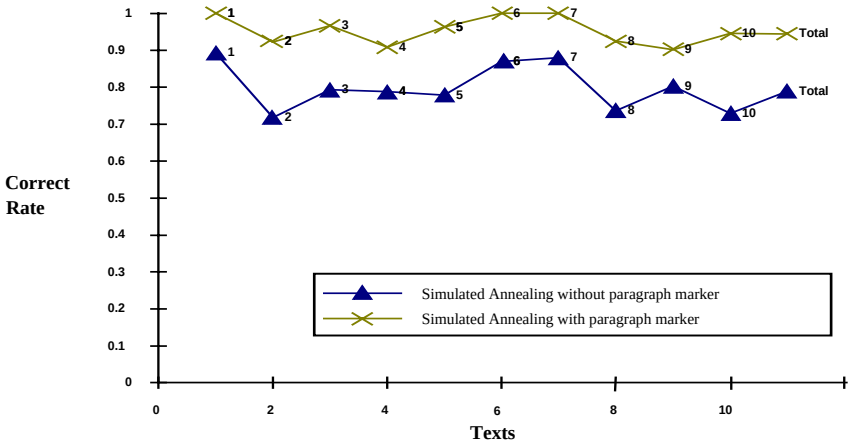

Figure 1. Comparison of Alignment Performance

The correct rates without and with using paragraph markers are 78.9% and 94.4%, respectively. The latter result (94.4%) is better than the bound correct rate (91.8%) mentioned before. It shows that those difficult bead types are resolved in our approach. Comparing Tables 5 and 6, we conclude that when the paragraph markers are used, the performance increases significantly. Fig. 1 shows the significance of paragraph markers. In other words, if an alignment algorithm could use any reliable anchor points in the texts, the performance will increase sharply.

In fact, the performance of alignment is dependent on the nature of the texts. When aligning a noisy texts without reliable anchor points, we will definitely do a bad job. However, the simulated annealing approach could reduce the risk, and the performance will keep over 78% in our experiment.

6. Conclusion

A new criterion to aligning texts is proposed in this paper. The criterion is based on an observation that the source texts and the target texts should share the same concepts, entities, ideas, and events. Sentence length (no matter word-based or character-based) (Brown, et al., 1991; Gale and Church, 1991a), is not so critical on languages across different language families. Translation-based criterion (Kay, 1991; Chen, 1993) is very useful, but it is also very complex. Surely, to decide word correspondences is difficult. Our criterion provides a tradeoff between the length-based criterion and the translation-based criterion. The clues of critical POSes are partially syntactic and partially statistical.

The performance of simulated annealing approach to alignment is 94% in our experiment, if we use the paragraph markers. Without paragraph marker, the value drops to 78%. Generally speaking, it works well for languages across different language families.

The main contribution of this work is to provide an alignment algorithm for aligning oriental languages with occidental languages. The future work should focus on the large experiment, normalizing the weight of critical POSes and other search techniques.

Acknowledgements

Research on this paper was partially supported by National Science Council grant NSC84-0408-E002-005.

References

Aarts, E. and J. Korst (1989). *Simulated Annealing and Boltzmann Machines*, John Wiley & Sons.

Behavior Design Corporation (1992). *The BDC Chinese Tagged Corpus*, Taiwan, R.O.C.

Brown, P. *et al*. (1989). "A Statistical Approach to Language Translation." *Proceedings of COLING*, pp. 71-76.

Brown, P. *et al.* (1991). "Aligning Sentences in Parallel Corpora." *Proceedings of 29th Annual Meeting of the ACL*, pp. 169-176.

Chen, S. (1993). "Aligning Sentences in Bilingual Corpora Using Lexical Information." *Proceedings of 31st Annual Meeting of the ACL*, pp. 9-16.

Dagan, I., A. Itai and U. Schwall (1991). "Two Languages are More Informative Than One." *Proceedings of 29th Annual Meeting of the ACL*, pp. 130-137.

Eijk, P. (1993). "Automating the Acquisition of Bilingual Terminology." *Proceedings of the Sixth Conference of the European Chapter of the ACL*, pp. 113-119.

Gale, W. and K. Church (1991a). "Identifying Word Correspondences in Parallel Texts." *Proceedings of Fourth DARPA Workshop on Speech and Natural Language*, pp. 152-157.

Gale, W. and K. Church (1991b). "A Program for Aligning Sentences in Bilingual Corpora." *Proceedings of 29th Annual Meeting of the ACL*, pp. 177-184.

Gale, W., K. Church and S. Yarowsky (1992). "Using Bilingual Materials to Develop Word Sense Disambiguation Methods." *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 101-112.

Johansson, S. (1986). *The Tagged LOB Corpus: Users' Manual*, Bergen: Norwegian Computing Centre for the Humanities.

Kay, M. (1991). "Text-Translation Alignment." *Conference Handbook of ACH/ALLC'91: Making Connections*, Tempe, Arizona, p. 257.

Sells, P. (1985). *Lectures on Contemporary Syntactic Theories*, Lecture Notes, No. 3, CSLI.

Sumita, E., H. Iida and H. Kohyama (1990). "Experiments and Prospects of Example-Based Machine Translation." *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 203-212.

Yang, Y. (1981). *Researches on Punctuation Marks*, Tien-Chien Publishing Company, Hong Kong.

Appendix

The testing corpus, CE_01 to CE_10, are selected from Sinorama Magazine (光華雜誌). The details of these texts are listed in the following.

CE_01: 劉蘊芳 (Liu, Yung-fang/tr. by Phil Newell), "學校是可以這樣辦的 (Is This Any Way to Run a School ?)," *光華雜誌* (*Sinorama Magazine*), Jan. 1991, pp. 108-111.

CE_02: 趙淑俠 (Chao, Shu-hsia/tr. by Peter Eberly), "僑社一員 (A Member of the Overseas Chinese Community)," *光華雜誌* (*Sinorama Magazine*), Mar. 1991, pp. 110-111.

CE_03: 張靜茹 (Chang, Chin-ju/tr. by Phil Newell), "「做人」的煩惱 -- 不孕 (It's Hard to Conceive--Infertility in Taiwan)," *光華雜誌* (*Sinorama Magazine*), May 1991, pp. 22-23.

CE_04: 琦君 (Ch'i Chiin/tr. by Peter Eberly), "閒情 (Idle Thoughts)," *光華雜誌* (*Sinorama Magazine*), May 1991, pp. 94-95.

CE_05: 陳雅玲 (Chen, Elaine/tr. by Peter Eberly), "加州理工的「精兵」生涯 (Caltech's "Crack-Troop" Way of Life)," *光華雜誌* (*Sinorama Magazine*), June 1991, pp. 124-125.

CE_06: 李光真 (Li, Laura/tr. by Christopher Hughes), "韓國也有「難念的經」(A Curse on Both Our Houses)," *光華雜誌* (*Sinorama Magazine*), Sept. 1991, pp. 40-41.

CE_07: 魏宏晉 (Wei, Hung-chin/tr. by Christopher Hughes), "「霍亂」現在進行式(Cholera-Present Progressive Tense)," *光華雜誌* (*Sinorama Magazine*), Nov. 1991, p. 47.

CE_08: 魏宏晉(Wei, Hung-chin/tr. by Phil Newell), "小心！電腦病毒就在你身邊 (Computer Viruses - It Can Happen to You)," *光華雜誌* (*Sinorama Magazine*), April 1992, pp. 34-38.

CE_09: 滕淑芬 (Teng, Sue-feng), "重新發現亞洲--「躍升中的亞洲經濟」國際會議記要 (Rediscovering Asia - The International Conference on "The Asian Regional Economy")," *光華雜誌* (*Sinorama Magazine*), June 1992, pp. 22-26.

CE_10: 林靜芸 (Lin, Ching-yun/tr. by Jonathan Barnard), "書評--哀悼乳房 (Book Review -- Mourning My Breast)," *光華雜誌* (*Sinorama Magazine*), Feb. 1993, pp. 90-92.