Aligning Parallel Chinese-English Texts Using Multiple Clues

Hsin-Hsi Chen and Yeong-Yui Wu

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.
E-mail: hh_chen@csie.ntu.edu.tw
Fax: 886-2-3628167

Abstract

Parallel texts bring much linguistic information, so that they can be applied to word-sense disambiguation, extraction of translation templates, automatic translation of noun compounds, construction of bilingual dictionary, and so on. For these works, sentence alignment is indispensable. Previous literatures seldom touch on the parallel texts from different language families, e.g., Chinese and English. This paper considers multiple clues such as critical punctuation marks, numbers, personal names, machine-made lexicon and man-made lexicon. The former four clues contribute different scores, and the latter one provides constraints for alignment. Three language models are presented, and their effects are discussed. Under the experiments on texts selected from Sinorama, a magazine published in Chinese and in English monthly by Government Information Office of R.O.C., the recall and the precision are 93.6% and 92.1%, respectively.

1. Introduction

Bilingual corpus has attracted the attention of many researchers because two languages can contribute more information than one (Dagan, *et al*., 1991). Such information can be used in word-sense disambiguation (Gale, *et al.*, 1992), extraction of translation templates (Kaji & Kida, 1992), automatic translation of noun compounds (Rachow, *et al*., 1992), construction of bilingual dictionary or terminological bank (Eijk, 1993), and so on. For the above applications, to align bilingual texts is an indispensable task. The tasks on alignment have three different levels such as paragraph, sentence and word. Several different approaches, e.g., length-based alignment (Brown, *et al*., 1991; Church, 1993; Wu, 1994), lexicon-based alignment (Chen, 1993; Kay & Rosenschein, 1993), part-of-speech-based alignment (Chen & Chen, 1994), and so on, are proposed. Most of these works focus on the materials selected from Hansards Corpus (an English-French corpus). The programs (Brown, *et al.*, 1991; Chen, 1993; Church, 1993; Kay & Rosenschein, 1993) show good performance on these testing data. A few papers (Wu, 1994; Chen & Chen, 1994) discuss the sentence alignment on different language families, e.g., parallel Chinese-English texts. Wu (1994) counts each Chinese character as having length 2 under Big 5 encoding system, and employs a specific bilingual correspondence lexicon to enhance the performance. The length criterion of this approach is problematic. Different encoding systems, e.g., EUC (Extended Unix Code) system, may have different character length. Chen & Chen (1994) apply critical part-of-speeches to aligning parallel texts. The result is promising. This paper proposes a different approach. We use multiple clues such as critical punctuation marks, numbers, personal names, machine-made lexicon and man-made lexicon. Three language models are presented, and their effects on sentence alignments are discussed.

2. Working Bilingual Corpus

Before discussion, the working corpus called *NTU bilingual corpus* is introduced. It is used to extract machine-made lexicon, and serve as testing data for different language models. Texts in this corpus are selected from Sinorama, a magazine published in Chinese and in English monthly by Government Information Office of R.O.C. It contains twenty Chinese texts and twenty English texts. Appendix A lists the sources of NTU bilingual corpus. Table 1 shows the statistics of this corpus. There are 738 Chinese sentences and 1002 English sentences. In total, there are 21493 Chinese words and 21118

English words. On the average, a Chinese sentence

Table 1. Statistics of NTU Bilingual Corpus

text codes	Total Chinese sentences/ words	Total English sentences/ words
T01	22/641	28/581
T02	32/1279	55/1339
T03	47/825	42/919
T04	37/846	50/832
T05	27/837	41/779
T06	25/822	35/819
T07	26/779	36/745
T08	56/1711	85/1825
T09	67/1583	88/1824
T10	37/573	63/545
T11	49/1035	47/1033
T12	51/1427	66/1487
T13	63/2086	92/2350
T14	26/680	27/705
T15	44/1419	64/1458
T16	24/684	30/623
T17	17/444	25/535
T18	24/743	35/757
T19	25/713	36/853
T20	39/971	57/1168
Sum	738/21493	1002/21118

consists of 29 words, and an English sentence consists of 21 words. In other words, Chinese sentences are longer than English ones. It is not strange that a Chinese sentence usually corresponds to more than one English sentence. Table 2 demonstrates the alignment patterns in NTU bilingual corpus. A pattern m-n denotes *m* Chinese sentences corresponds to *n* English sentences. Because Chinese and English belong to different language families, the sentence patterns are more complex than those in occidental family. The 1-1 pattern still forms the majority. However, there are many other more complex cases such as 1-3, 1-4, 1-5, 1-6, 2-3, 2-4, 3-1, *etc*. That makes sentence alignment much harder. Till now, most works only consider five patterns -say, 1-0, 0-1, 1-1, 1-2 and 2-1. The proportion of these patterns in our working corpus is 88%. Intuitively, if we apply the conventional approaches to this corpus, the upper bound of performance is 88%.

3. Clues for Sentence Alignments

When a source sentence in one language is translated into a target sentence in another language, some words may disappear and some words are very likely to appear in the translation pairs called *sentence beads* (Brown, *et al*., 1991). Punctuation marks, numbers, and personal names are very critical tokens, so that they often occur in parallel sentences. Table 3 demonstrates the statistics got from NTU bilingual corpus. The notation m-n denotes m and n punctuation marks occur in Chinese and English parts of a sentence bead, respectively. Total patterns and the proportions are shown in the second and the third rows. The scores shown in the last row will be discussed latter.

3.1 Punctuation Marks

The proportion of English and Chinese sentences that contain question marks at the same time is 86%. In a parallel text, if a sentence contains a question mark, it is likely to be aligned to a sentence that also contains this mark. The following demonstrates an example:

(C1) 這兩所學校究竟憑著什麼魅力風迷世界各地?

(E1) What makes these two schools so famous? (quoted from T01)

The following also shows a counter example to this criterion:

(C2) 韓國在國際經濟中所展現的野心及成績，已引起我們企業界「十年後，韓國會不會成爲第二個日本？」的憂慮。

(E2) The ambitions and achievement of Korea in the world economy have already led Taiwan's businessmen to worry that in ten years' time she might become a second Japan. (quoted from T06)

Similarly, the proportions of the bilingual sentences that have exclamation marks (quotation marks) at the same time are 62% (69%). The following show some (counter) examples for exclamation marks (the first two examples) and quotation marks (the second two examples):

Table 2. The Alignment Patterns in NTU Bilingual Corpus

Text Codes	1-0	1-1	1-2	1-3	1-4	1-5	1-6	2-1	2-2	2-3	2-4	3-1
T01	0	16	6	0	0	0	0	0	0	0	0	0
T02	0	15	11	2	1	1	0	0	0	1	0	0
T03	17	16	11	1	0	0	0	1	0	0	0	0
T04	0	19	6	4	0	0	0	2	1	1	0	0
T05	1	18	3	3	4	0	0	0	0	0	0	0
T06	0	11	9	1	0	0	0	1	1	0	0	0
T07	0	14	9	1	0	0	0	1	0	0	0	0
T08	0	30	16	5	1	0	0	1	0	1	0	0
T09	1	43	14	2	1	0	0	1	1	0	1	0
T10	0	16	14	1	1	0	1	0	1	0	1	0
T11	14	23	6	6	6	0	0	0	6	6	0	0
T12	0	37	13	1	0	0	0	0	0	0	0	0
T13	0	36	16	7	1	0	0	1	1	0	0	0
T14	0	22	2	0	0	0	0	1	0	0	0	0
T15	0	22	7	4	2	0	0	1	0	2	0	1
T16	0	19	2	0	0	1	0	0	1	0	0	0
T17	0	10	2	3	0	0	0	0	1	0	0	0
T18	0	10	4	2	2	0	0	3	0	0	0	0
T19	0	10	7	2	0	0	0	0	1	0	0	0
T20	0	19	14	2	0	0	0	0	2	0	0	0
Sum	33	406	172	42	11	2	1	13	13	6	2	1
Proportion (%)	4.70	57.8	24.5	6.00	1.57	0.28	0.14	1.85	1.85	0.86	0.28	0.24

(C3) 看他抱著不得不送到收容所的小狗小貓，輕聲細語地對牠說：「喊我哥哥呀，哥哥好捨不得你啊！」我心中真是不忍。

(E3) Seeing him hug a puppy or kitten that had to be sent to the pound as he quietly told it, "You're my little brother. I hate to give you up!" was truly unbearable. (quoted from T04)

(C4) 如此一來，不但原本可觀的報酬岌岌可危，還可能賠上違約金！

(E4) Not only is his anticipated reward now endangered, he may even have to pay a fine for violating the contract. (quoted from T08)

(C5) 與森林小學一樣，夏山學校和巴氏學園都是「體制外」的兒童教育機構，而且學生人數都很少，全校只有五、六十人。

(E5) Like the Forest Elementary School, both the Summerhill and Pashih schools are "outside the system." Both have few students- only 50 or 60. (quoted from T1)

(C6) 爲了提供孩子鮮活的知識，他甚至請來「農夫老師」給他們上了一節活生生的農事教學課。

(E6) To give them knowledge of the vitality of life, he even arranged to have a farmer come and teach about farming. (quoted from T01)

Table 3. The Statistics of Critical Punctuation Marks

(1) Question Marks:

patterns	0-1	1-0	1-1	1-2	2-0	2-1
# of patterns	2	5	45	1	1	3
proportion(%)	3.5	8.7	78.9	1.7	1.7	5.2
score	0	0	7.89	0.17	0	0.52

(2) Exclamation Marks:

patterns	0-1	1-0	1-1	1-2
# of patterns	1	6	9	1
proportion(%)	5.9	3.5	56.2	5.9
score	0	0	5.62	0.59

(3) Quotation Marks:

patterns	0-2	1-1	2-0	2-2	4-2	4-4
# of patterns	8	12	59	110	10	17
proportion(%)	3.7	5.6	2.7	50.9	4.6	7.9
score	0	0.56	0	5.09	0.46	0.79

3.2 Numbers

Number is the second critical token. The following demonstrates a typical example:

(C7) 夏山學校是英國人尼爾於一九二一年在蘇格蘭所成立。

(E7) The Summerhill School was founded in Scotland by A. S. Neill in 1921. (quoted from T01)

However, several problems may be introduced when number is considered as a clue. First, the indefinite article "a" is often used with a singular noun in English. However, there is no similar rule in Chinese. That is, "a" is not always translated into the Chinese word "一". Take (C6) and (E6) as an example. "A former" is translated into "農夫老師" rather than "一個農夫老師". To avoid confusion, Chinese word "一" and English word "a" are neglected. Second, Chinese number and English one may have different values, but denote the same thing. For example, "在民國七十八年" corresponds to "in 1989". Here, the Chinese word "七十八" and the English word "1989" denote the same year. It can be disambiguated by the keyword "民國". Third, the use of the caesura sign, which is a Chinese specific punctuation mark, may result in different representation of numbers. For example, "五、六十" in (C5) corresponds to "50 or 60" in (E5). On the surface, "五" (five) is translated into "50".

3.3 Personal Names

Personal name is another clue. The following demonstrates a typical example:

(C8) 我們特別請到作家齊邦媛作評，並訪問作者。

(E8) We have specially invited Chi Pang-yuan to review the book and have interviewed the author. (quoted from T11)

The problem is: personal name is often an unknown word. How to identify them is a difficult problem (Mani, *et al*., 1993; Lee, *et al.,* 1994). This problem decreases the power of the clue.

3.4 Lexicons

A Chinese sentence and an English one are likely to be matched if some words in one language are the translation of another language. Two kinds of lexicons -say, machine-made and man-made, are available. The former is trained from bilingual corpus, and the latter

Table 4. A Bad Example for Performance Evaluation

RBS	(1,1), (2,3), (1,0), (1,2)
IRBS	**(1,1)**, **(3,4)**, **(4,4)**, **(5,6)**
CBS	(1,1), (0,1), (1,0), (0,1), (0,1), (1,0), (1,0), (0,1), (0,1), (1,0)
ICBS	**(1,1)**, (1,2), (2,2), (2,3), (2,4), **(3,4)**, **(4,4)**, (4,5), (4,6), **(5,6)**
Recall	4/4=1.0=100%

is a conventional electronic dictionary. In our work, we manually align 3/4 of NTU bilingual corpus. Then, the co-occurrences of English words and Chinese words are considered. They are used to compute the similarity of distribution. If the co-occurrence frequency of two parallel words is high, their distribution is similar, so that the two words tend to be the translation of each other. To avoid interference, two monolingual corpora, i.e., NTU segmented corpus and Brown corpus, are used to get the higher frequent words and the lower frequent words. These words are filtered out when the machine-made lexicon is set up.

4. Evaluation Model

Although a lot of experimental results have been reported with high correct rates, only the paper (Chen & Chen, 1994) touches on how to evaluate the performance when answers are in terms of bead sequence. We use bead sequence to obtain the alignment performance (recall). An alignment consists of a sequence of beads. The correct sequence of beads is called *Real Bead Sequence* (RBS). In contrast, the sequence of beads generated by the alignment algorithm is called *Computed Bead Sequence* (CBS). To evaluate the performance, *Incremental Bead Sequence* (IBS) is defined further as follows:

> *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.

The performance (recall) is measured by the following formula:

$$\text{performance} = \frac{\text{number of common beads between } IRBS \text{ and } ICBS}{\text{number of beads in } IRBS}$$

In fact, recall cannot express the real performance of sentence alignment. Assume the output of an alignment algorithm is shown in Table 4. The recall will be very high (100%). However, the result is very bad. To avoid this problem, precision is defined too.

$$\text{precison} = \frac{\text{number of common beads between } IRBS \text{ and } ICBS}{\text{number of beads in } ICBS}$$

In this way, the precision of the result shown in Table 4 is 40%.

5. Baseline Model

The first alignment program consists of four main modules: Chinese segmentation system, English morphological analyzer, token identifier, and alignment module. Because a Chinese sentence is composed of a string of characters without any boundaries, and the basic meaningful unit is a word instead of a character, Chinese segmentation system first finds the words in each Chinese sentence. English morphological analyzer converts nouns and verbs into their root forms. Token identifier identifies the critical tokens like punctuation marks, numbers and names in the texts. Finally, the alignment module aligns the parallel texts by dynamic programming. Here, only the alignment module is discussed in detail.

5.1 Score Assignments

The alignment algorithm gives a score to each proposed sentence bead, and finds the alignment with the largest sum of scores. Consider the following sentence bead:

C_{i1} C_{i2} C_{i3} C_{i4} . . . . . C_{in}
E_{j1} E_{j2} E_{j3} E_{j4} . . . . . E_{jm}

C_{ip} is the p-th token in the i-th Chinese sentence, and Ejq is the q-th token in the j-th English sentence. The score of each sentence bead is determined by four factors: content words, punctuation marks, numbers and names. The score assigned by content word is denoted by *C(i,j)*, and is computed in the following way:

$$C(i,j) = \frac{\sum_{q=1}^{m}\sum_{p=1}^{n} T(i_p, j_q)}{\sum_{q=1}^{m}\sum_{p=1}^{n} U(i_p, j_q)}$$

In the baseline model, the machine-made lexicon is used. That is, $T(i_p,j_q)$ is the value looked up from the training table for pair (i_p,j_q). $U(i_p,j_q)$ equals to 1 if pair (i_p,j_q) can be found in the training table; otherwise $U(i_p,j_q)$ equals to 0.

Table 3 shows different punctuation marks have different power. The proportion of each pattern is normalized into a value (score) between 0 and 10. Score 0 is assigned to a pattern without critical punctuation marks, i.e., m-0 and 0-m. Furthermore, a

bead with a punctuation pattern not listed in this table has the smallest value in the corresponding punctuation table, e.g., 0.17 for question mark. The score contributed by critical punctuation marks is denoted by *Pu(i,j)*.

Pu(i,j) = P(i,j)+Q(i,j)+R(i,j)

Here, *P(i,j)*, *Q(i,j)* and *R(i,j)* are the scores contributed by exclamation marks, question marks and quotation marks, separately.

Number is very critical in sentence alignment. The score assigned by number is denoted by *Nu(i,j)*.

*Nu(i,j) = 10*s*

where *s* is the total equal-number tokens.

In contrast to number, personal name has weaker power. This is because it is often an unknown word. How to identify it is a serious problem. Our work (Lee, *et al*., 1994) on the identification of Chinese personal names is adopted. It has precision 88.04% and recall 92.56%. The problem to identify the English names in the English texts is: the first character of English name has upper case, but not every word satisfying this condition is a name. We adopt a rule: transliterated English name has a form of "Family-Name (initial-capital), Given-Name (initial-capital)". Because the rule is rough, only the numbers of possible personal names instead of the exact matching of personal names are used. Assume there are *c* names in a Chinese sentence C_i and *e* names in an English sentence E_j. The score contributed by names in C_i and E_j is denoted by *Na(i,j)*.

if $(c = e \neq 0)$ then *Na(i,j)* = 5

else *Na(i,j) = 4/(abs(c-e))*

Consider a general sentence bead. Assume Chinese part consists of m+1 sentences, i.e., C_{i-m}, $C_{i-(m-1)}$, ..., C_i, and the corresponding English part has n+1 sentences, i.e., E_{j-n}, $E_{j-(n-1)}$, ..., E_j. The score of this sentence bead is *s(i,j;m,n)*.

s(i,j;m,n) = C(i,j;m,n) + Pu(i,j;m,n) + Nu(i,j;m,n) + Na(i,j;m,n)

C, *Pu*, *Nu* and *Na* denote the scores contributed by content words, punctuation marks, numbers and names, respectively.

5.2 Alignment Algorithm

Table 2 shows there are various bead types in parallel Chinese-English texts. In general, if a paragraph has at most *n* sentences, and those (0-m) (or (m-0)) bead patterns are regarded as *m* occurrences of (0-1) (or (1-0)) patterns, there are n^2+2 possible bead patterns. Assume n=6. Thus, we have 38 possible bead patterns. The alignment algorithm is formulated recursively. Let C_i (i=1, ..., I) be Chinese sentences, and E_j (j=1, ..., J) be the corresponding English translations. Let *s* be the score described in the previous section, and *S(i,j)* be the maximum score among sentences C_1, C_2, ..., C_i and their translations E_1, E_2, ..., E_j under the maximum likelihood alignment. *S(i,j)* is computed recursively, where the recurrence maximizes over the 38 cases. *S(i,,j)* is calculated as follows with initial condition *S(i,j)* = 0.

S*(i,j)* = max(*S(i-m,j-n) + s(i,j;m,n)*)

Here, m $(0 \le m \le 6)$ and n $(0 \le n \le 6)$ cannot be equal to 0 at the same time.

5.3 Experimental Results

The second column (Model 1) in Table 5 shows the experimental results of the baseline model. Because the training data is not large enough, the similarity distribution in machine-made lexicon is not critical. The performance does not meet our expectation.

6. Advanced Models

In this section, an English-Chinese dictionary is used as an aided resource. When the corresponding translations of words in English also appear in Chinese part, it is very likely to align these two sentences together.

6.1 Model 2

Because an English word may have more than one translation, and the corresponding translation may appear more than once in the Chinese part, only words that appear in only one sentence are considered as clues. These are called *critical words*. We can imagine there is a link between English and Chinese sentences when there are critical word pairs by dictionary lookup. The links can serve as extra constraints to the baseline model. Consider the following figure.

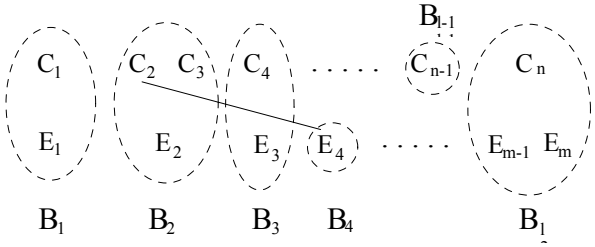

Table 5. Experimental Results

Text	Model 1	Model 2	Model 3

Codes	Recall	Precision	Recall	Precision	Recall	Precision
T01	0.95	0.95	1.00	1.00	1.00	1.00
T02	0.81	0.79	1.00	0.97	0.97	0.94
T03	0.83	0.83	0.90	0.90	0.90	0.90
T04	0.76	0.74	0.91	0.86	0.91	0.86
T05	0.96	1.00	0.92	1.00	0.96	0.96
T06	0.87	0.87	1.00	1.00	0.96	0.96
T07	0.92	0.92	0.92	0.92	0.92	0.92
T08	0.83	0.82	0.93	0.93	0.89	0.87
T09	0.73	0.73	0.89	0.89	0.92	0.91
T10	0.81	0.76	0.97	0.92	0.94	0.89
T11	0.79	0.77	0.94	0.91	0.94	0.91
T12	0.90	0.94	0.92	0.98	0.92	0.94
T13	0.84	0.85	0.80	0.82	0.92	0.93
T14	0.96	0.96	0.96	0.96	0.96	0.96
T15	0.95	0.90	0.95	0.90	0.97	0.93
T16	0.91	0.95	0.91	0.95	0.91	0.95
T17	0.94	0.94	0.94	0.88	1.00	0.94
T18	0.86	0.86	0.95	0.95	0.95	0.95
T19	0.86	0.83	0.91	0.83	1.00	0.92
T20	0.86	0.86	0.84	0.79	0.92	0.89
Average	0.854	0.849	0.918	0.910	0.936	0.921

The sentence beads B2 and B3 violate this constraint, so that they must be discarded. To decrease the cost of dictionary look-up, only the important content words, i.e., nouns and verbs, are considered. The score of a sentence bead in this model is shown as follows:

$$s(i,j;m,n) = Pu(i,j;m,n) + Nu(i,j;m,n) + Na(i,j;m,n)$$

The factor of machine-made lexicon is replaced by man-made lexicon. The third column (Model 2) in Table 5 shows the precision and the recall increase 6.1% and 6.4%, respectively. Some texts aligned by Model 2 have worse performance. There are two major reasons. First, the English-Chinese dictionary we used in the experiment has only 44962 entries. The critical content words may not be in this dictionary. Second, even they have the entries, the corresponding translation may not be the same as the words in the texts. Segmentation error is one of the reasons.

6.2 Model 3

The machine-made lexicon can be regarded as a specific resource trained from special corpus, and the man-made lexicon is regarded as a generic resource. Model 3 integrates all the clues from different sources. The total score $s(i,j;m,n)$ is shown as follows:

$$s(i,j;m,n) = C(i,j;m,n) + Pu(i,j;m,n) + Nu(i,j;m,n) + Na(i,j;m,n)$$

Man-made lexicon provides the constraints as before. The last column in Table 5 shows the experimental results of Model 3. Because not all the critical words can be found in the man-made lexicon, and they may be trained from bilingual corpus beforehand, the precision and the recall increase 1.1% and 1.8% compared to Model 2. Appendix B lists the alignment result of text T18 for reference.

7. Concluding Remarks

This paper applies different clues to aligning parallel Chinese and English texts. According to the critical degrees, punctuation marks, numbers, personal names and machine-made lexicon contribute different scores. Dynamic programming technique is used to find the alignment with the maximal scores under the constraints provided by man-made lexicon. Such an approach integrates knowledge from various sources. Under the experiments on very difficult texts, parts of which are translated word-by-word and parts of which are translated by meaning, the recall and the precision are 93.6% and 92.1%, respectively.

References

Brown, P.F.; Lai, J.C. and Mercer, R.L. (1991) "Aligning Sentences in Parallel Corpora."

Proceedings of 29th Annual Meeting of the ACL, pp. 169-176.

Chen, K.-H. and Chen, H.-H. (1994) "A Part-of-Speech-Based Alignment Algorithm." *Proceedings of COLING-94*, pp. 166-171.

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

Church, K.W. (1993) "Char-align: A Program for Aligning Parallel Texts at the Character Level." *Proceedings of 31st Annual Meeting of the ACL*, pp. 1-8.

Dagan, I.; Itai, A. and Schwall, U. (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 Sixth Conference of the EACL*, pp. 113-119.

Gale, W.; Church, K. and Yarowsky, S. (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.

Kaji, H.; Kida, Y. and Morimoto, Y. (1992) "Learning Translation Templates from Bilingual Text." *Proceedings of COLING-92*, pp. 672-678.

Kay, M. and Roscheisen, M. (1993) "Text-Translation Alignment." *Computational Linguistics*, 19(1), pp. 121-142.

Lee, J.-C.; Lee, Y.-S. and Chen, H.-H. (1994) "Identification of Personal Names in Large-Scale Texts." *Proceedings of ROC Computational Linguistics Conference*, pp. 203-222.

Mani, I.; Macmillan, T.R.; Luperfoy, S.; Lusher, E. and Laskowski, S. (1993) "Identifying Unknown Proper Names in Newswire Text." *Proceedings of Workshop for the Acquisition of Lexical Knowledge from Text*, pp. 44-54.

Rackow, U.; Dagan, I. and Schwall, U. (1992) "Automatic Translation of Noun Compounds." *Proceedings of COLING-92*, pp. 1249-1253.

Wu, D. (1994) "Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria." *Proceedings of 32nd Annual Meeting of ACL*, pp. 80-87.

Appendix A. Working Bilingual Corpus

The working bilingual corpus contains 40 texts selected from Sinorama Magazine (光華雜誌). They belong to different domains. The details of these texts are listed in the following.

T01: 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.

T02: 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.

T03: 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.

T04: 琦君 (Ch'i Chiin/tr. by Peter Eberly), "閒情 (Idle Thoughts)," *光華雜誌* (*Sinorama Magazine*), May 1991, pp. 94-95.

T05: 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.

T06: 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.

T07: 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.

T08: 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.

T09: 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.

T10: 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.

T11: 齊邦媛 (Chi, Pang-yuan/tr. by Jonathan Barnard), "少年大頭春的生活週記 (Weekly Report of Young Big Head Chun)," *光華雜誌* (*Sinorama Magazine*), Jan. 1993, pp. 82-84.

T12: 張瓊方 (Chang, Chung-fang/tr. by Phil Newell), "「方寸」之間的樂趣 (Fun by the Square Inch)," *光華雜誌* (*Sinorama Magazine*), July 1992, pp. 84-87.

T13: 滕淑芬 (Teng, Sue-feng/tr. by Christopher Hughes), "洋學位，土麵包--就業市場看留學情結 (Foreign Study, Domestic Bread-Overseas Students in the Job Market)," *光華雜誌* (*Sinorama Magazine*), Oct. 1992, pp. 13-16.

T14: 蔡文婷 (Ventine Tsai/tr. by Andrew Morton), "「在野」戲團生命力 (Vitality of Drama Troupes "Out in the Boondocks")," *光華雜誌* (*Sinorama Magazine*), March 1992, pp. 37.

T15: 陳淑美 (Chen, Jackie /tr. by Peter Eberly), "三角習題：男女工作平等法 (The Equal Employment Act: Relief for Working Moms (and Dads))," *光華雜誌* (*Sinorama Magazine*), March 1992, pp. 24-25.

T16: 陳雅玲 (Chen, Elaine/tr. by Jonathan Barnard), "家庭價值的新趨勢 (The Family Comes Back in Style)," *光華雜誌* (*Sinorama Magazine*), Jan. 1993, pp. 35.

T17: 張瓊方 (Chang, Chung-fang/tr. by Brad Baudler), "不一樣的比賽 (A Different Kind of Competition)," *光華雜誌* (*Sinorama Magazine*), Feb. 1992, pp. 46.

T18: 張靜茹 (Chang, Chin-ju/tr. by Jonathan Barnard), "科技的「幸福指標」 (A Scientific "Happiness Index")," *光華雜誌* (*Sinorama Magazine*), May. 1992, pp.32-33.

T19: 張靜茹 (Chang, Chin-ju/tr. by Christopher Hughes), "污染農地何去何從？ (Polluted Farmland -- From Where to Where)," *光華雜誌* (*Sinorama Magazine*), Nov. 1992, pp. 53-54.

T20: 李光真 (Li, Laura/tr. by Phil Newell), "誰來監督國會？ (Who Will Watch the Parliament?)," *光華雜誌* (*Sinorama Magazine*), Nov. 1992, pp. 7-8.

Appendix B. Sample Result

The following shows the results of text T18. Each bead is marked on a pattern type or an error symbol (***).

1:1 科技 指標 既然 不 等於 幸福 指標 ； 那 怎麼樣 才能 使 科技 的 發展 ， 爲 人們 帶來 較 多 的 「 幸福 」 ？
Since indexes of scientific development are not a gauge of happiness, how indeed can scientific development be engineered to add most to the quality of people's lives?

1:1 過去 廿 年來 的 投入 ， 使 我國 科技 人才 漸 出 ， 建立 了 自己 的 科技 社 群 ； 研 發 品質 日 漸 改善 ； 發展 科技 以 改善 民生 的 目的 ， 也 初步 成 。
Over the last 20 years, Taiwan has gradually established a pool of scientific talent; the quality of its research has been growing by the day; and the first steps have been taken to develop science for the purpose of improving people's lives.

1:1 就 在 對 繁榮 前景 ， 寄予 無限 期望 的 同時 ， 我們 也 無可 避免地 爲 科技 帶來 的 高 成長 付出 代價 ： 環境污染 、 自然 資源 急速 耗竭 ， 甚至 人性 的 庸俗 化 … … 。
At a time of limitless expectations about a prosperous future, we have no way to avoid the heavy costs of technology: environmental pollution, the rapid depletion of natural resources, and even human vulgarity.

1:2 臺大 大氣 科學 系 系主任 林和 指出 ， 從 十七 世紀 科學界 建立 了 唯一 可以 單向 累積 成果 的 科學 知識 ， 科技 進步 之 輪 已是 無法 阻擋 。
Lin Ho, the head of the atmospheric science department at National Taiwan University, points out that since the 17th century the scientific world has chugged down a single track on which scientific knowledge has accumulated.
The wheels of technological progress are already impossible to stop from turning.

1:3 回頭 路 已失 ， 我們 更 無法 一聲令下 叫 各 國 停止 科技 競爭 ； 科技 即使 能 全面 撤退 ， 也 無法 解決 問題 。
There is no way to turn back. We have even less chance of asking all countries to call a halt to technological competition. And even if technology could be disentangled from all aspects of life, we still couldn't get to the root of the problem.

1:2 我們 不能 脫離 現實 ， 我們 仍 需要 科技 來 防止 衰退 造成 的 貧窮 ， 「 但 在 生活 已無 匱乏 之際 ， 也 應 多 做 另 一 種 價值 思考 。」
We cannot sever ourselves from the realities of the present day. We still need technology to prevent decline that would lead to poverty, "But when we are already free from want, we should adopt another value system."

1:3 林和 以爲 ， 現代 科技 與 人們 生活 密切 相關 ， 科學界 更 需 考慮 公眾 責任 ， 應 不時 回頭 思考 ： 我們 的 科技 發展 是否 一定 要 盲目 的 和 其他 國家 一較高下 ？
Lin Ho believes that modern technology is closely connected to people's lives.
These in the scientific community ought to consider their public responsibilities.
Every once in a while, they ought to turn around and consider such questions as these: Must we always blindly compare our level of technological development to those of other countries?

1:1 科技 發展 在 這 個 社會 是爲 誰 服務 ？
Whom is technological development serving in this society?

1:1 同樣 的 資源 是否 應 另 有所 用 ？
Should we use our resources in other ways?

***「 人們 只有 在 科技 價值 上做 抉擇 ， 」 林和 舉例 說 ： 「 我們 是 要 軍事 科技 ？ 或 綠色 科技 ？ 投入 高溫 超導 ， 是爲 了 改良 軍火 武器 ？
"Our decisions about technology have to be based on our values," says Lin Ho. "For example, do we want military technology or green technology?

*** 或 它 可以 在 資訊 工業 上 有 大 突破 ， 以 大量 代替 紙張 ， 讓 森林 與其 相關 的 生命 ， 都 得以 間接 少 受 傷害 ？」
Shall we develop high-temperature superconductivity for use in military weapons or for major breakthroughs in the information industry, reducing the need for paper so that the forests and other related forms of life can indirectly receive less harm?"

1:1 科技役於人、或人爲科技所役，選擇權仍在人手上。
Whether technology serves man or man serves technology is still man's choice.

1:1 過去蘇聯全力投入國防科技的抉擇，無疑是給人們的當頭棒喝：它證明了，光是科技內部盲目的昇華，並不能保證人們的幸福，反而連國家最基礎的生活都無法維持。
The example of the former Soviet Union has woken people up: it proves that blindly seeking after technological advancement can neither guarantee people's happiness nor provides support for the most basic of human needs.

1:1 在這個時代，我們應更認真的從各種可能的觀點了解科技，把它和人生各個層面連在一起觀察。
In this age, we should work hard to obtain a multi-layered understanding of technology from many perspectives that is combined with an understanding of humanity.

1:2「人與科技」一書編者孫志文指出，在科技思考已侵入了工程、農業、商業、傳播系統、教育制度的今天，「唯有充滿自我省悟的批判，才是投入科技不致迷失方向的要件。」
Sun Chih-wen, the author of People and Technology, points out that technological thought had already taken over engineering, agriculture, commerce, communications systems, and educational structures. "Only by possessing critical self awareness can one not lose one's way in the midst of technology."

1:1 否則科技知識累積愈豐，也將愈快忘記科技知識存在的意義。
Otherwise, as technological knowledge grows ever more abundant, one will increasingly forget why technological knowledge exists.

1:4 科技快速發展，造成高度工業化國家與未開發國家間生活水準的差距，使得佔有十二億人口的工業先進國家，消耗了世界上八分之七的資源，而擁有四十億人口的所謂「第三世界」，許多地區卻呈饑餓狀態，這正是科技人員應該思考的課題。
The rapid development of technology has resulted in a gap between the living standards of the highly industrialized countries and the under-developed countries. As a result, the 1.2 billion people in the advanced industrial countries consume seven-eighths of the resources. And the four billion residents of the "Third World" are faced with hunger in many areas. This is also a topic that those involved in science and technology ought to consider.

1:4 因此，在討論國防、交通等課題時，都不能只由科技的選擇上著眼，比如交通發展必需注意到未來的車輛及飛機對於都市、環境等所可能造成的影響；在我們研究、擬定未來科技發展計劃時，也要由全球的觀點考量，舉凡環境的維護、開發中國家技術的支援、世界性能源的合理使用等均不可或缺。
Hence, in discussing national security, transportation and other such issues, one cannot merely turn one's attention to technological choices. For example, in developing transportation we must pay attention to how future cars and airplanes will affect cities and the environment. In researching technology development plans for the future we must consider things from a global perspective. Such issues as environmental protection, technological support for the developing countries and the rational use of global resources cannot be overlooked.

2:1 交通大學教務長，前國家科學委員會副主委鄧啓福以爲，我們需要以國際胸襟，對地球負一份責任。例如參與世界性的「全球變遷」計劃，對全球環境的破壞付出關注。
Den Chi-fu, the dean of studies of National Chiao Tung University and the former vice chairman of National Science Council, says we need to be internationally minded and take responsibility for the earth -- for example, by participating in the worldwide "Global Change" plan, which is concerned about damage to the global environment.

1:2 我們也應把自己的科學成果，視爲全人類共同的福祉，以全世界爲對象，開放我們建立起來的同步輻射中心、高速電腦中心等國家實驗室，特別是提供機會給尚無能力維持研究設施的國家。
We also ought to use our technological accomplishments to benefit all people.
We should open up such national laboratories as the Synchrotron Radiation Research Center an the High Speed Computer Center for the entire world, in particular providing opportunities for countries that cannot yet support their own research facilities.

1:1 在鄧啓福看來，這樣的胸懷，也才是科技力最高層次的展現。
In the view of Den Chi-fu, this kind of mind-set represents the highest level of technological power.