Computational Linguistics and Chinese Language Processing
Vol. 10, No. 1, March 2005, pp. 95-122

Aligning Parallel Bilingual Corpora Statistically with Punctuation Criteria

Thomas C. Chuang* **and Kevin C. Yeh+**

Abstract

We present a new approach to aligning sentences in bilingual parallel corpora based on punctuation, especially for English and Chinese. Although the length-based approach produces high accuracy rates of sentence alignment for *clean* parallel corpora written in two Western languages, such as French-English or German-English, it does not work as well for parallel corpora that are noisy or written in two disparate languages such as Chinese-English. It is possible to use cognates on top of the length-based approach to increase the alignment accuracy. However, cognates do not exist between two disparate languages, which limit the applicability of the cognate-based approach. In this paper, we examine the feasibility of exploiting the statistically ordered matching of punctuation marks in two languages to achieve high accuracy sentence alignment. We have experimented with an implementation of the proposed method on parallel corpora, the Chinese-English Sinorama Magazine Corpus and Scientific American Magazine articles, with satisfactory results. Compared with the length-based method, the proposed method exhibits better precision rates based on our experimental reuslts. Highly promising improvement was observed when both the punctuation-based and length-based methods were adopted within a common statistical framework. We also demonstrate that the method can be applied to other language pairs, such as English-Japanese, with minimal additional effort.

Keywords: Sentence Alignment, Cognate Alignment, Machine Translation

* Department of Computer Science, Vanung University, No. 1 Van-Nung Road, Chung-Li Tao-Yuan, Taiwan, ROC
E-mail: tomchuang@cc.vit.edu.tw
+ Department of Computer Science, National Tsing Hua University, 101, Kuangfu Road, Hsinchu, 300, Taiwan, ROC

1. Introduction

Bilingual corpora are very important for building natural language processing systems [Moore 2002; Gey *et al.* 2002], including data-driven machine translation [Dolan *et al.* 2002], computer-assisted revision of translations [Jutras 2000], and cross-language information retrieval [Chen and Gey 2001]. In order to develop NLP systems, it is useful to align bilingual corpora at the sentence level with very high precision [Moore 2002; Chuang *et al.* 2002, Kueng and Su 2002]. With aligned sentences, further analysis such as phrase and word alignment analysis [Ker and Chang 1997; Melamed 1997], bilingual terminology [Déjean *et al.* 2002] and collocation [Wu 200] extraction analysis can be performed. Yang, C, Li, K. [2003] proposed an alignment method for bilingual title pairs on the Web for automatic generation of bilingual parallel corpora. The hybrid dictionary approach [Collier *et al.* 1998], text-based alignment [Kay and Röscheisen 1993], part of speech-based alignment [Chen and Chen 1994], and the lexical method [Chen 1993] are other examples of sentence alignment methods. While these methods presume little or no prior knowledge of source and target languages, they are relatively complex and require significant amounts of parallel text and language resources.

Much work reported in the computational linguistics literature has focused on aligning English-French and English-German sentences. While the length-based approach [Gale and Church 1993; Brown *et al.* 1991] to sentence alignment produces surprisingly good results for French and English with success rates well over 96%, it does not work well for the alignment of English and Chinese sentences. Simard, Foster, and Isabelle [1992] proposed using cognates on top of the length-based approach to improve the alignment accuracy. They use an operational definition of cognates, which include digits, alphanumerical symbols, punctuation, and alphabetical words. Several other measures of cognateness have also been suggested [Melamed 1999; Danielsson and Muhlenbock 2000; Ribeiro *et al.* 2001], but none of them are sufficiently reliable, and all of them are tailored to close Western language pairs.

Simard, Foster, and Isabelle [1992] pointed out that cognates in two close languages, such as English and French, can be used to measure the likelihood of mutual translation. Those cognates include alphabetic words, numeric expressions, and punctuation that are almost identical and readily recognizable by computers. However, for disparate language pairs, such as Chinese and English, that lack a shared Roman alphabet, it is not possible to rely on such cognates to achieve high-precision sentence alignment of noisy parallel corpora.

Research on sentence alignment of English and Chinese texts [Wu 1994], indicates that the lengths of English and Chinese texts are not as highly correlated as they are with French and English, leading to lower precision rates (86.4-95.2%) for length based aligners. A comparison of the correlation between German-English and Chinese-English bilingual corpora is depicted in Figure 1, where 138 German- English and 151 Chinese-English aligned

sentences are analyzed. The correlations are 0.99 and 0.95 for the German-English and Chinese-English cases, respectively. The expected ratios and the corresponding standard deviations are (0.92, 0.1124) and (4.614, 0.84) for the German-English and Chinese-English cases, respectively.

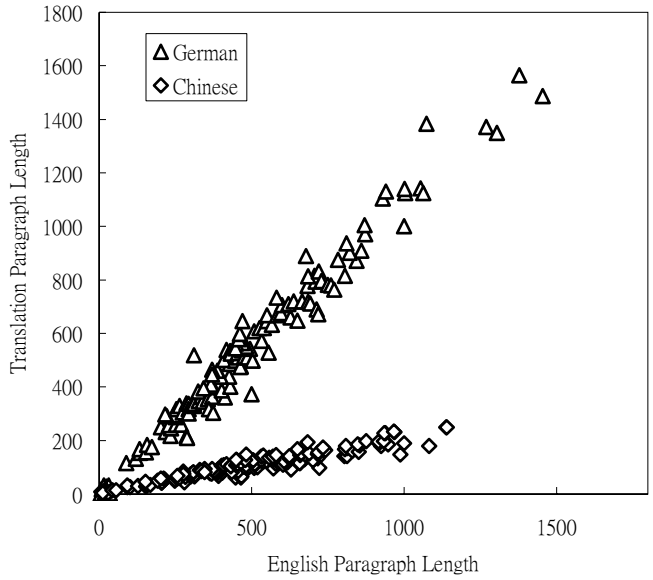

Figure 1. The relationships between German-English [Gale and Church 1993] and Chinese-English bilingual paragraph lengths [Chuang et al. 2002]. The correlations are 0.99 and 0.95 for the German-English and Chinese-English cases, respectively. The expected ratios and the corresponding standard deviations are (0.92, 0.1124) and (4.614, 0.84) for the German-English and Chinese-English cases, respectively.

Furthermore, for English-Chinese alignment tasks, no orthographic, phonetic, or semantic cognates, that are readily recognizable by computer exist. Therefore, the inexpensive cognate-based approach is not applicable to Chinese-English tasks. We are thus motivated to find alternative evidence that two blocks of texts are mutual translations. It turns out that punctuation can be telling evidence, if we do more than hard matching of punctuation and take into consideration intrinsic sequencing and the statistical distribution of punctuation in ordered comparison. What is attractive about this approach is that it easily leads to sub-sentential alignment, which has been shown to be useful for statistical machine translation.

Section 2 of this paper, we provide some information about the similarity in the use of punctuation in Chinese and English literature and also the differences. Our conclusion is that using punctuation as cognates to align disparate parallel texts will fail to provide adequate alignment results. Section 3, we define a punctuation compatibility factor as an indicator of mutual translation. A translation model that employs a punctuation probability function is proposed. In Section 4, we present experiments based on our novel approach of using the statistical properties of punctuation in parallel texts being analyzed to perform bilingual sentence alignment. We demonstrate that one can use punctuation alone to develop a high-precision sentence alignment program for distant parallel texts like those in Chinese-English corpora. Additionally, we examine the performance of sentence alignment by using punctuation in combination with length. In Section 5, we demonstrate that the proposed method is a very cost effective approach that can be effectively applied to other disparate bilingual languages like English-Japanese without *a priori* language knowledge of them. A brief conclusion is provided in Section 6.

2. Punctuation across languages

According to the Longman Dictionary of Applied-Linguistics [Richards *et al.* 1985], a *cognate* is "a word in one language which is similar in form and meaning to a word in another language because both languages are related." Although the ways in which different languages around the world use punctuation vary, symbols such as commas and full stops are used in most languages to demarcate writing, while question and exclamation marks are used to show interrogation and emphasis. However, these forms of punctuation can often look different or be used in different ways.

The traditional Chinese writing system does not have punctuation, and it is up to the reader to demarcate the text while reading. With the influx of Western culture in the eighteenth century, punctuation systems similar to the one used with Roman script was adopted in China and Japan. The punctuation includes the period, comma, colon, dash, etc. Although most of those forms of punctuation look similar to Roman ones, they are usually coded as double-bytes and tend to be used differently. The full stop in Chinese and Japanese is a small empty circle, quite different in appearance from the Roman period. Quotes are also very different, shaped like a Greek letter Γ, upright or upside down. There are forms of punctuation that have no counterparts in Roman text. For instance, "、" is the pause symbol, which is used somewhat like the comma but only when separating items in a list. On the other hand, there are several uses of the Roman comma which do not occur in Chinese texts. A few examples are given below:

(Parenthetical expressions)

(1e) Evolution, as far as we know, doesn't work this way.

(1c) 我們所知道的進化論不是如此的。

(Appositives)

(2e) His father, Tom, is a well-known scholar.

(2c) 他的父親湯姆是一位有名的學者。

Yang [1981] described more punctuation marks in Chinese used in various ways that are similar or dissimilar to English punctuation. In summary, although Simard et al. [1992] considered the various forms of punctuation in English and French to be cognates, in general, punctuation forms are not cognates for many other language pairs.

In both Chinese and English texts, the average ratio of the punctuation count to the total number of tokens available is low (less than 15%). But punctuation provides valid additional evidence, which can help one achieve a high degree of alignment precision. Our method can easily be generalized to other language pairs since minimal a priori linguistic knowledge is required.

3. Punctuation and Sentence Alignment

3.1 Punctuation Marks in English and Chinese

In this section, we will describe how punctuation in two languages can be used to measure the likelihood of mutual translation in sentence alignment. We will use an example in the following to illustrate the method. A formal description also follows:

Example 3 shows a Chinese sentence and its translation counterpart of two English sentences in a parallel corpus.

(3c) 逐漸的，打鼓不再能滿足他，「打鼓原是我最喜歡的，後來卻變成邊打邊睡，一個月六萬元的死工作」，薛岳表示。

(3e) Over time, drums could no longer satisfy him. "Drumming was at first the thing I loved most, but later it became half drumming, half sleeping, just a job for NT$60,000 a month," says Simon.

If we keep punctuations in the above examples in the original order and strip everything else out, we have ten pieces of punctuation from the English part (3e) and eight from the Mandarin part (3c) as follows:

(4c)	，	，	「	，		，	」	，	。
(4e)	,	.	"	,	,	,	,	"	.

They can be arranged into different match types as shown below.

Match type	**(4c)**	**(4e)**
1-1	，	,
1-1	，	.
1-1	「	"
1-1	，	,
0-1		,
1-1	，	,
2-2	」 ，	, "
1-1	。	.

Figure 2. The correspondence between two punctuation strings

There are several frequently used punctuation forms in Chinese text that are not available in English text, for example, the punctuation forms "、" and "。". These punctuation forms often correspond to the English punctuation forms "," and ".", respectively. It is not difficult to see that the two punctuation strings above match up quite nicely, indicating that the corresponding texts are mutual translations. Roughly, the first two commas in Chinese correspond to the first two English punctuation marks (comma and period), while the Chinese open quote in the third position corresponds to the English open quote also in the third position. The two Chinese commas inside the quotes correspond to two of the four commas within the quotes in English. The two consecutive marks (」，) correspond to (,"), forming a 2-2 match. These correspondences can be unraveled via a dynamic programming procedure, much like sentence alignment. See Figure 2 for more details.

It is apparent that the punctuation in the two strings match up very consistently, and that the matching is somewhat continuous with respect to the alignment of regular words surrounding the punctuation (see the double-lined links in Figure 3 for details). Therefore, the example gives a convincing indication that the correspondence between punctuation across two languages can provide telling evidence that two texts are mutual translations.

3.2 Punctuation marks as Good Indicators of Mutual Translation

Based on our initial observation, the portion of the identifiable punctuation matches between two parallel texts in Chinese and English is over 50%. Examining Figure 2, we can identify institutively the matches between the Chinese punctuation and the equivalent English punctuation marks: (「) corresponds to (“), etc. This implies that although direct match information is useful, there is still a large discrepancy in the punctuation mappings between Chinese and English. We, therefore, define here a punctuation compatibility factor that can be used to further analyze the relationship between the punctuation found in parallel texts. The punctuation compatibility factor as an indicator of mutual translation is defined as

$$\gamma = \frac{c}{\max(n, m)}, \tag{1}$$

where γ = the punctuation compatibility factor,
c = the number of direct punctuation matches,
n = the number of Chinese punctuation marks,
m = the number of English punctuation marks.

We took aligned English-Chinese sentences that had the same punctuation count (which is the denominator of Equation 1), take ten for example, in order to determine how well punctuation works as an indicator of mutual translation of English and Chinese sentences. We also took the same English sentences and matched them up with randomly selected Chinese sentences to calculate the compatibility of punctuation marks in unrelated texts.

The results obtained indicated that the average compatibility of pairs of sentences, which were mutual translations, was about 0.67 (with a standard deviation of 0.170), while the average compatibility of random pairs of bilingual sentences was 0.34 (with a standard deviation of 0.167).

Figure 3. English punctuation across aligned sentences

Figures 4 through 6 show the compatibility results based on punctuation counts of eight, ten and twelve respectively. These graphs were constructed by analyzing around 50,000 aligned sentences found in the Sinorama Magazine (1990-2000). 521, 259, and 143 sentences were selected to obtain values of n and m equal to 8, 10, and 12, respectively. The solid lines simply connect data points for easier observation.

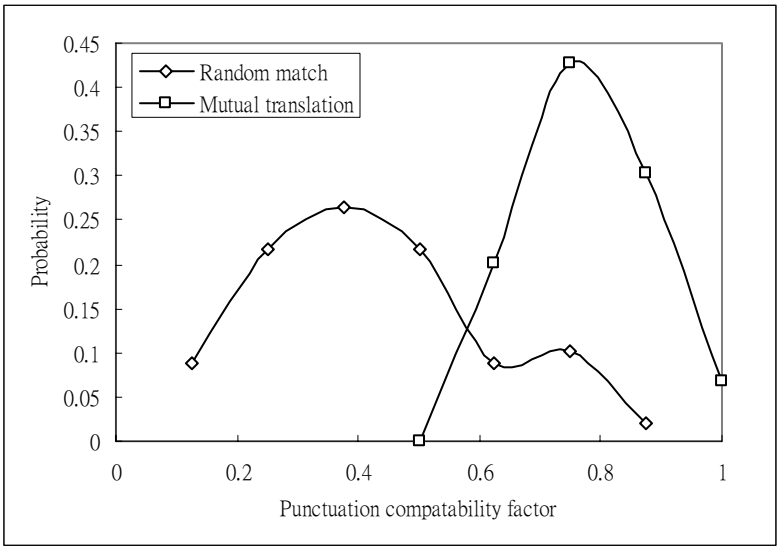

Figure 4. Compatibility of translation pairs vs. random pairs with n=m=8

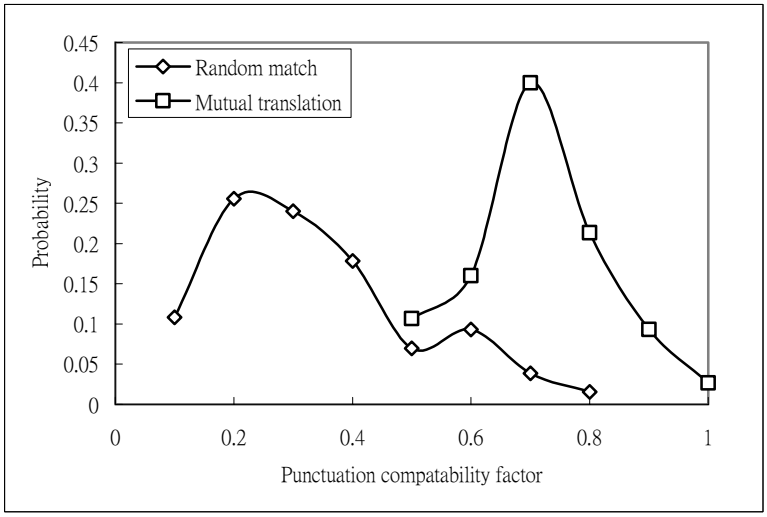

Figure 5. Compatibility of translation pairs vs. random pairs with n=m= 10

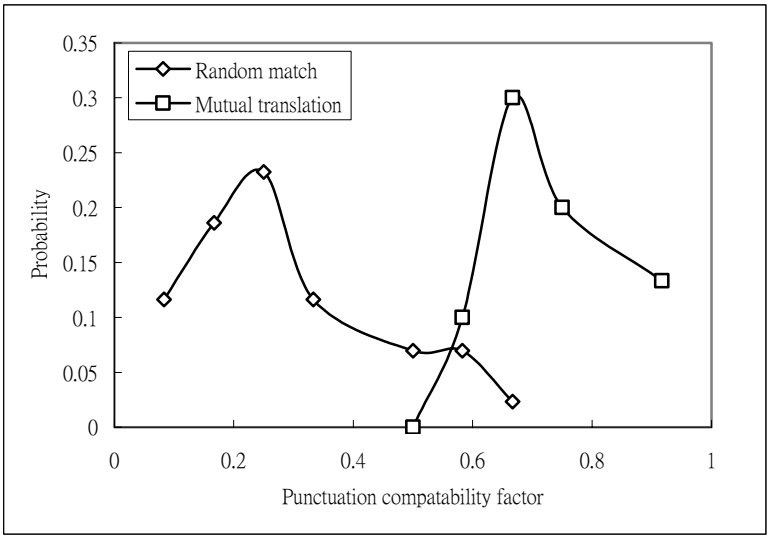

Figure 6. Compatibility of translation pairs vs. random pairs with n=m = 12

Intuitively, as the number of punctuation marks increases, the reliability of the compatibility function does also. Overall, if the punctuation marks are softly matched in ordered comparison across the two languages, they indeed provide useful information for effective sentence alignment. Analyzing the Sinorama corpus, we found that the percentage of matched sentences having the same number of punctuation marks was 21.42%. We selected and analyzed aligned sentences having different numbers of punctuation marks to get more insight into the distinction between matched and random sentences. The analysis also helped us to determine the proper use of the binomial distribution function for sentence alignment. Sentences with eight, ten, and twelve punctuation marks were arbitrarily chosen for analysis. It appears that the distinction between mutual translations and unrelated texts indeed becomes more prominent for sentences that have larger numbers of punctuation marks.

3.3 Punctuation Alignment Model

Instead of one-to-one hard matching of punctuation marks in parallel texts as used in the cognate approach of Simard et al. (1992), we allow no match and one-to-several matching of punctuation matches. Our model of the probability of punctuation alignment is very similar to the word alignment model proposed by Brown et al. (1991). In order to perform soft matching of punctuation, we define the probability that a sequence of punctuation marks $CP_i = Cp_1 Cp_2 Cp_3 \cdots Cp_i$ in a sentence in the source language (L1) translates into a sequence of punctuation marks $EP_i = Ep_1 Ep_2 Ep_3 \cdots Ep_i$ in a sentence in the target language (L2) as $P(EP_i, CP_j)$. We choose the punctuation alignment that maximizes the probability overall

possible alignments, given a pair of punctuation sequences corresponding to a pair of parallel sentences, i.e.,

$$\arg\max_{A} P(A|CP_i, EP_j) \quad ,$$

where A is a punctuation alignment. Assuming that the probabilities of the individually aligned punctuation pairs are independent and applying the Bayes' rule, we can make the following approximation:

$$P(EP_i, CP_j) \approx \prod_{A} P(Cp_k, Ep_k) \cdot P(|Cp_k|, |Ep_k|) \qquad (2)$$

where $|Cp_k|$ and $|Ep_k|$ = the number of punctuation marks in CP_i and EP_j ,respectively, which ranges from 0 to 2,

$P(Cp_k, Ep_k)$ = the probability of translating Cp_k into Ep_k , and

$P(|Cp_k|, |Ep_k|)$ = the probability of translating $|Cp_k|$ punctuations in L1 into $|Ep_k|$ punctuation in L2.

We observe that in most cases, the links of punctuation do not cross each other, much like the situation with sentence alignment. Therefore, it is possible to use the dynamic programming procedure to softly match punctuation across languages.

In order to explore the relationship between punctuation in pairs of Chinese and English sentences that are mutual translations, we selected a small set of manually aligned texts and investigated the characteristics and the statistics associated with the punctuation. Information from around 500 manually analyzed sentences was then used as the initial parameters to bootstrap a larger corpus. An unsupervised EM algorithm and dynamic programming were used to optimize the punctuation correspondence between a text and its translation counterpart. The steps in the standard EM algorithm which we used included initializing model parameters with manually analyzed punctuation matching probabilities, assigning probabilities to missing punctuation, estimating model parameters from completed data, and iterating the process until convergence was reached. The EM algorithm converged quickly after the second iteration of training.

We observed that, in most cases, the links of punctuation did not cross each other, much like the situation with sentence alignment. Therefore, we were motivated to use the dynamic programming procedure to *softly match* punctuation across the languages by finding the Viterbi path using the punctuation translation function $P(Cp_k, Ep_k)$ and fertility function $P(|Cp_k|, |Ep_k|)$. The translation probability functions corresponding to 1-1, 2-2, 1-0, and 0-1 English-Chinese punctuation matches are shown in Tables 1 to 4, respectively. It should be noted that the calculated probability was the conditional probability of each punctuation mark, therefore, the sum of the probability in each table does not equal to one.

Table 1. The frequency counts and the conditional probabilities of 1-1 English-Chinese punctuation matches, sorted according to count

E_p	C_p	Match type	Count	Prob.
,	，	1-1	541	0.809874
.	。	1-1	336	0.657528
"	﹁	1-1	131	0.34203
.	，	1-1	113	0.221133
"	﹂	1-1	112	0.292423
"	「	1-1	65	0.16971
"	」	1-1	59	0.154044
,	、	1-1	56	0.083832
,	。	1-1	41	0.061377
!	！	1-1	38	0.883508
.	…	1-1	30	0.058708
?	。	1-1	17	0.447277
:	，	1-1	12	0.666302
;	、	1-1	11	0.422925
,	「	1-1	10	0.01497
?	？	1-1	9	0.236794
.	、	1-1	7	0.013698
"	，	1-1	7	0.018276
;	，	1-1	7	0.269134
.	；	1-1	6	0.011742
"	：	1-1	6	0.015666
,	：	1-1	5	0.007485
?	，	1-1	5	0.131552
,	；	1-1	4	0.005988
.	—	1-1	4	0.007828
:	：	1-1	4	0.222101
;	。	1-1	4	0.153791
)	）	1-1	4	0.997159
,	·	1-1	3	0.004491
.	·	1-1	3	0.005871
.	﹁	1-1	3	0.005871
!	。	1-1	3	0.069751
?	—	1-1	3	0.078931

Table 2. The frequency counts and conditional probabilities of 2-2 English-Chinese punctuation matches, sorted according to count

E_p	C_p	Match Type	Count	Prob.
,"	」’	2-2	6	0.956403
."	」’	2-2	3	0.916449
?"	———	2-2	2	0.611063
.!	…！	2-2	1	0.785235
!"	⌒⌣	2-2	1	0.785235
?"	」°	2-2	1	0.305531
??	———	2-2	1	0.785235

Table 3. The frequency counts and conditional probabilities of 1-0 English-Chinese punctuation matches, sorted according to count

E_p	C_p	Match Type	Count	Prob.
,		1-0	106	0.3655
.		1-0	66	0.2276
"		1-0	59	0.2034
)		1-0	23	0.0793
(		1-0	20	0.0691
:		1-0	7	0.0241
?		1-0	5	0.0172

Table 4. The frequency counts and conditional probabilities of 0-1 English-Chinese punctuation matches, sorted according to count

E_p	C_p	Match Type	Count	Prob.
	’	0-1	229	0.389455
	—	0-1	58	0.098639
	°	0-1	52	0.088435
	」	0-1	50	0.085034
	、	0-1	45	0.076531
	「	0-1	41	0.069728
	…	0-1	39	0.066326
	？	0-1	14	0.02381
	└—	0-1	14	0.02381
	：	0-1	9	0.015306
	—┐	0-1	7	0.011905

The punctuation match types (also known as the fertility functions) obtained through training are summarized in Table 5. Notice that the counts shown in the table are not integers because the results of EM training were adjusted using the Good Turing Smoothing Method to improve them.

Table 5. The punctuation fertility functions

Punctuation Match Type	Count	Probability
0-1	588.0005	0.225027
1-0	286.001	0.109452
1-1	1698.076	0.649852
1-2	2.466198	0.000944
2-1	0.965034	0.000369
2-2	37.19216	0.014233

3.4 Punctuation-based Sentence Alignment Model

Unlike the method Simard et al. [1992] used to handle cognates, we model the *compatibility* of punctuation across two languages using the Binomial distribution. Each punctuation mark appearing in one language either has one to three punctuation counterparts across translation or does not. For each punctuation mark, the probability of it having a translation counterpart is independent with a fixed value of *p*. Our approach differs from Simard's in the following interesting ways:

1. We use the Binomial distribution, while Simard et al. used a likelihood ratio.
2. We go beyond hard matching of punctuation marks between parallel texts. We allow a punctuation mark in one language to match up with a number of compatible punctuation marks in another. The compatibility model is similar in structure to the lexical translation probability proposed by Brown et al. [1991].
3. We take into consideration the intrinsic sequencing of punctuation marks in an ordered comparison. A flexible and ordered comparison of punctuation is carried out via dynamic programming.

Following Gale and Church [1993], we employ the Bayes Theorem to estimate the likelihood of aligning two text blocks E and C by calculating P(*E*, *C*|*match*) P(*match*).We adopt the same dynamic programming method, but use punctuation marks to measure the likelihood of mutual translation instead of lengths. The proposed sentence alignment method is based on a model in which each punctuation mark in L1 is responsible for generating a number of punctuation marks with a given matching probability in L2.

We define the probability of mutual translation for a given alignment pattern P(A| C,E) as follows: Given two blocks of text *E* and *C*, we first strip off non-punctuation therein and

determine the maximum number of punctuation marks *n* in either *E* or *C*.

We employ punctuation-based sentence alignment, which maximizes the probability of overall possible alignment, given a pair of parallel texts, i.e.,

$$\arg\max_{A} P(A \mid C, E),$$

where A is an alignment and C and E are the source and target texts, respectively.

A further approximation encapsulates the dependence of a single parameter b, which is a function of CP and EP:

$$P(A|Ci,Ej) = P(A|b(CP,EP)) .$$

Since it is easier to estimate the distribution for the inverted form, we apply Bayes' Rule to further simplify the calculation:

$$P(A|b) = P(b|A)P(A)/P(b) ,$$

where P(b) is a normalizing constant that can be ignored during minimization. P (A) is the match type, and its values are shown in Table 6. We use a binominal distribution to estimate P(b):

$$P(A \mid C, E) \approx \prod_{A} P(A \mid C_i, E_j)$$

$$\approx \prod_{k=1}^{t} P(A_k) \cdot \binom{n_k}{r_k} P(Cp_k, Ep_k)^{r_k} (1 - P(Cp_k, Ep_k))^{n_k - r_k} \quad , \tag{3}$$

where n_k = the maximum number of punctuation marks in either the English text or the Chinese text in the k[th] sentence to be aligned;

r_k = the number of compatible punctuation marks in ordered comparison;

$P(Cp_k, Ep_k)$ = the probability of the existence of a compatible punctuation mark in both languages;

$P(A_k)$ = the match type probability of aligning $E_{i,k}$ and $C_{j,k}$;

t = the total number of sentences that are aligned.

From the data, we have found that about 66% of the time, a sentence in one language matches exactly one sentence in the other language (1-1). Three additional possibilities should be also considered: 1-0 (including 0-1), and many-1 (including 1-many). Chinese-English parallel corpora are quite noisy, reflecting from wider possibilities of the match types. Here, we used the same probabilistic figures as proposed by Chuang and Chang [2002]. Table 6 shows all eight possibilities used in our implementation.

Table 6. The sentence alignment match type probability P(A)

P(A)	1-1	1-0, 0-1	1-2	2-1	2-2	1-3	3-1
Chinese-English	0.64	0.0056	0.017	0.25	-	0.056	-

3.5 A Hybrid Punctuation-based and Length-Based Sentence Alignment Model

The length-based sentence alignment criterion involves a length-related probability distribution function $P(\delta \mid match)$, where δ is a function of the sentence length of the source language l_c and the sentence length of the target language l_e, or $\delta = \delta(l_c, l_e)$. Since the sentence lengths of the bilingual parallel texts of interest are highly correlated, $P(\delta \mid match)$ can be estimated using the Gaussian assumption following Gale and Church [1993]. Incorporating both the length-based and punctuation-based criteria, we can modify equation (3) as follows:

$$P(A \mid C, E) \approx \prod_{k=1}^{t} P(A) \cdot P(\delta \mid match) \cdot \binom{n_k}{r_k} P(CP_k, EP_k)(1 - P(CP_k, EP_k))^{n_k - r_k} \quad . \tag{4}$$

The same dynamic programming optimization can then be used. Again, the computation and memory costs are very low when both the length-based and punctuation-based criteria are employed. The average slopes of l_c and l_e, and the associated standard deviations are estimated in an adaptive manner for each corpus being evaluated [Chuang *et al*. 2002].

4. Experiments and evaluation

To explore the relationship between the punctuation marks in pairs of Chinese and English sentences that are mutual translations, we prepared a small set of 200 pairs of sentences aligned at the sentence and punctuation levels. We then investigated the characteristics of and the statistics associated with the punctuation marks. We derived estimates of the punctuation translation probabilities and fertility probabilities from the small set of hand-tagged data. This seed information was then used to train the punctuation translation model on a larger corpus via the EM algorithm. The probability of a punctuation mark having a translation counterpart was estimated as p = 0.670 with a standard deviation 0.170. For random pairs of bilingual sentences, p = 0.340, with a standard deviation 0.167. There appears to be marked differences between the two distributions, indicating that, indeed, soft and ordered comparison of punctuation marks across languages provide useful information for effective sentence alignment.

In order to assess the performance of punctuation-based sentence alignment, we randomly selected five bilingual articles from the Sinorama Magazine Corpus and Scientific American (US and Taiwan editions), and several chapters from the novel Harry Potter. These were subjected to an implementation of the proposed method. Some experimental results are shown in appendices A and B.

It should be noted that in Appendix A, the first English sentence and the first Chinese sentence are both title sentences, and that they are aligned based on the carriage return

deliminater, even though no punctuation marks are found in the English sentence. We found that, in general, there were more periods in the English text than that in the Chinese text for a given bilingual corpus, especially in the case of a text translated from Chinese into English. As an example, 112 periods were found in a Chinese article [Sinorama 1988], whereas only 180 periods were found in the corresponding English translation. This phenomenon can be further seen by examining the punctuation distribution. The relationship between the number of sentences and the number of punctuation marks per sentence for the English-Chinese corpus was determined by analyzing 6,103 articles, including around 130,000 sentences, from Sinorama Magazine between 1976 and 2000. 1,138,447 punctuation marks were found in the English corpus and 2,056,675 punctuation marks in the Chinese corpus. Apparently, punctuation marks are used more sparingly in Chinese sentences. As shown in Figure 7, there were two punctuation marks in most of the English sentences and three in most of the Chinese sentences. The figure also shows that a few long sentences had close to one hundred punctuation marks, but these were unuseful.

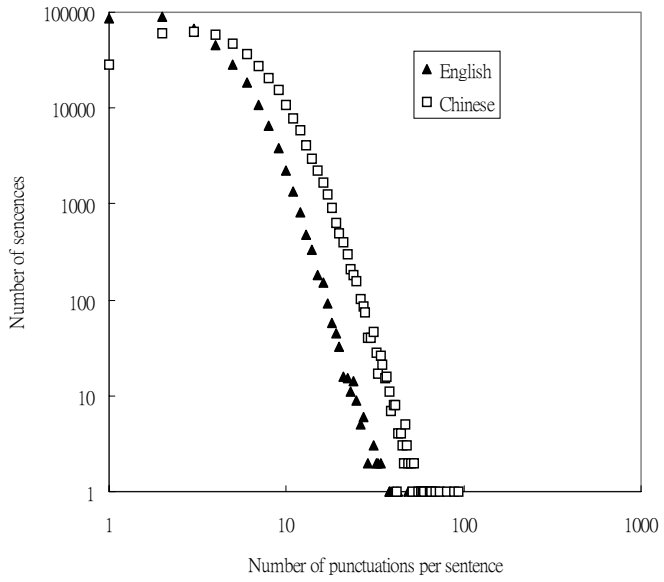

Figure 7. The punctuation distribution for a bilingual corpus

This observation prompted us to establish a special rule that the combination of a comma and an open quote in a Chinese sentence should be considered as being equivalent to a full stop. Applying this rule, we found that the sentence count increased from 112 to 126 for the Chinese text mentioned in the above example. This empirical rule helped to improve the

precision of sentence alignment. These special cases can be found in both Appendixes A and B.

The precision rate of the length-based approach [Gale and Church 1993] is shown in Table 7 as a baseline for comparison. The precision rate is defined as the ratio between the number of correctly matched sentences in the system output and the number of matched sentences generated from the system output. The large variation observed in the alignment precision is primary due to the disparity in the lengths and match types. The experimental results obtained with the punctuation-based approach and the combination approaches are shown in Tables 8 and 9. Overall, the punctuation-based approach outperformed the length-based approach, reducing the error rates consistently, and the improvement could exceed 50% at times.

Table7. Baseline sentence alignment performance achieved using the length-based approach

Articles	No. of Chinese Sentences	No. of Errors	Percentage (%)
World in a box*	75	7	90.7
What clones*	77	12	84.4
New University**	319	24	92.5
Book I-2 ***	439	16	96.4
Book II-8 ***	633	19	97.0

* Scientific American

** Sinorama

*** Harry Potter

Table 8. Performance evaluation using punctuations

Article	Baseline	Precision	Improvement
World in a box*	91.5	98.8	7.3
What clones*	86.5	96.6	10.1
New University**	93.0	95.3	2.3
Book I-2 ***	96.5	98.9	2.4
Book II-8 ***	97.1	98.0	0.9

Table 9. Performance evaluation by combining length and punctuation information

Article	Baseline	Precision	Improvement
World in a box*	91.5	100.0	8.5
What clones*	86.5	97.8	11.2
New University**	93.0	93.9	0.9
Book I-2 ***	96.5	96.7	0.2
Book II-8 ***	97.1	98.2	1.1

Additionally, we evaluated our method on a larger corpus the Scientific American Corpus. We used all of the English and Chinese articles from January 2003 to December 2003. There were 67 articles, 1523 English sentences, and 1599 Chinese sentences. Every article included both an English text and its corresponding Chinese text. The punctuation-based sentence alignment method achieved alignment precision rates of over 93%. Inferior performance was achieved when the hybrid punctuation and length-based method was used as compared with the punctuation-based method alone, as shown by the results listed in Tables 8 and 9. This phenomenon may be attributed to the strong dependence of the length-based method on the average length of the sentences being analyzed. Apparently, length-based methods do not perform well in the case of a corpus that is composed of shorter sentences. Therefore, a length-based method may achieve poorer performance when it is combined with a punctuation-based method. Consequently, caution should be exercised in interpreting these precision rates.

Our approach has been proven to be effective, and it has been used to construct a concordancer system called **TotallRecall** [Wu *et al.* 2003] in a Computer Assisted Language Learning (CALL) project. Based on the results of our experiments, it was also possible to speed up the corpus annotation and distribution efforts made by the Association for Computational Linguistics and Chinese Language Processing.

5. Discussion

We achieved a striking improvement over the length-based baseline for bilingual text alignment when punctuation was used alone or in combination with lexical information. Combining punctuation and length information, we could get slightly better overall performance. However, the improvement was not entirely consistent. Thus we need to experiment on a longer parallel text in order to be more certain about it.

Although word alignment links cross each other quite often, punctuation links do not. It appears that we can obtain sub-sentential alignment at the clause and phrase levels from the alignment of punctuation. For instance, after we align the punctuation in examples (3c) and

(3e), we can extract the following finer-grained bilingual analyses:

(5c) 逐漸的
(5e) Over time
(6c) 打鼓不再能滿足他
(6e) drums could no longer satisfy him
(7c) 打鼓原是我最喜歡的
(7e) Drumming was at first the thing I loved most
(8c) 後來卻變成邊打邊睡
(8e) but later it became half drumming, half sleeping
(9c) 一個月六萬元的死工作
(9e) just a job for NT$60,000 a month
(10c) 薛岳表示
(10e) says Simon

We have hand-coded a small English-Japanese punctuation mapping table and converted our alignment program to handle alignment of Japanese and English texts. It appears that the adapted program works with performance comparable to that of the original one. An example of aligning English-Japanese parallel texts based on punctuation is shown in Appendix C. Our Japanese-English program is a very preliminary one. Further and more rigorous investigation is needed.

6. Conclusion and Future Work

We have developed a very effective sentence alignment method based on punctuation. The probability of the finding matches between different punctuation marks in source and target texts is calculated based on a large bilingual corpus. The punctuation-based measure of mutual translation can be modeled by the binomial distribution. We have implemented the proposed method on the parallel Chinese-English Sinorama Magazine Corpus. The experimental results show that the punctuation-based approach outperforms the length-based approach with precision rates exceeding 93%.

We have also demonstrated that the alignment method can be applied to other bilingual texts, without the need for *a priori* linguistic knowledge of the languages, like Japanese and English. This general approach has been found to be fast, easy to set up, and universal. We believe that this method can be easily applied to many different languages.

A number of interesting future directions for researches present themselves. First, punctuation alignment can be exploited to constrain word alignment and reduce error rates. Second, punctuation alignment makes possible a finer-grained level of bilingual analysis of sub-sentential alignment and can provide a strikingly more effective translation memory and bilingual concordance for more effective example-based machine translation (EBMT), computer assisted translation and language learning (CAT and CALL).

Acknowledgement

We are indebted to Dr. Jason Chang for helpful discussions and suggestions. We acknowledge the support for this study provided through grants by the National Science Council and Ministry of Education, Taiwan (NSC-2213-E-238-015, NSC 90-2411-H-007-033-MC and MOE EX-91-E-FA06-4-4), and by the MOEA under the Software Technology for Advanced Network Application Project of the Institute for Information Industry.

Appendix A

Some experimental results of sentence alignment based on length and punctuation are presented here. Shaded parts indicate imprecision in alignment results. We calculated the precision rates by dividing the number of un-shaded sentences (counting both English and Chinese sentences) by the total number of sentences proposed. Since we did not exclude aligned pairs using a threshold, the recall rate should be the same as the precision rate. The experimental results indicate that when non 1-1 matches next to each other tend to fail the length-based aligner. However, the punctuation-based aligner appears to handle such cases more successfully.

Sentence alignment based on length		
Type	English text	Chinese Text
11	Take note	「共筆」怪現象
12	Allowing education to be led by the market may also lead to deficiencies in teaching practices.	市場領導教育還可能引發教學上的弊病。台大法律系教授賀德芬說，對法律系學生來說，考上司法官、高考是最好的出路，
11	Professor He Te-fen of NTU's Department of Law say that for law students, the best opportunity for advancement is to pass the recruitment examinations for public prosecutors and judges, or the senior civil service exams.	「有些學生上課只想具體知道如何答考題，選課標準就是老師的教書方式是不是對考試有用。」
31	"In class, some students only want to learn specifically how to answer exam questions, and their choice of courses depends on whether the instructor's teaching method is helpful for passing the exams." Some instructors, seeing that some students do not take good notes, even designate one who does to give them to the others for reference. But this results in most of the students taking no notes at all, because after all they will get photocopies, paid for out of the class expenses fund.	甚至有老師因爲看學生的筆記記不好，指定做得好的同學給其他人參考，以提高系上的錄取率，結果變成學生也不做筆記了，反正有班費可以影印給大家，這個現象還有個名詞叫「共筆」。
21	Two years ago, the CER completed a "General Consultation Report on Educational Reform." One of its main proposals was that the past system of controlling the establishment, expansion and contraction of departments in higher education on the basis of estimates of personnel demand should be "relaxed."	兩年前行政院教改會完成「教育改革總諮議報告書」，建議的重點之一是，過去以人力需求的推估，管制高等教育科系的設立增減，應該「鬆綁」。
11	Education cannot be made merely to narrowly serve the economy.	教育不能「窄化」成只爲經濟服務，但現實的狀況是，
11	Yet in reality, "the reason most parents are willing	「大部分家長之所以肯花錢讓孩子來

	to pay to put their children through university is certainly not that they hope they will become passionate seekers after truth, but to enable them to find good careers," says Providence University president Li Chia-tung bluntly.	念大學，絕不是希望孩子以後熱衷於真理的追求，而是爲了使孩子將來能找到好職業，」靜宜大學校長李家同明白地說。
Sentence alignment based on punctuation		
11	Take note	「共筆」怪現象
11	Allowing education to be led by the market may also lead to deficiencies in teaching practices.	市場領導教育還可能引發教學上的弊病。
11	Professor He Te-fen of NTU's Department of Law say that for law students, the best opportunity for advancement is to pass the recruitment examinations for public prosecutors and judges, or the senior civil service exams.	台大法律系教授賀德芬說，對法律系學生來說，考上司法官、高考是最好的出路，
11	"In class, some students only want to learn specifically how to answer exam questions, and their choice of courses depends on whether the instructor's teaching method is helpful for passing the exams."	「有些學生上課只想具體知道如何答考題，選課標準就是老師的教書方式是不是對考試有用。」
21	Some instructors, seeing that some students do not take good notes, even designate one who does to give them to the others for reference. But this results in most of the students taking no notes at all, because after all they will get photocopies, paid for out of the class expenses fund.	甚至有老師因爲看學生的筆記記不好，指定做得好的同學給其他人參考，以提高系上的錄取率，結果變成學生也不做筆記了，反正有班費可以影印給大家，這個現象還有個名詞叫「共筆」。
21	Two years ago, the CER completed a "General Consultation Report on Educational Reform." One of its main proposals was that the past system of controlling the establishment, expansion and contraction of departments in higher education on the basis of estimates of personnel demand should be "relaxed."	兩年前行政院教改會完成「教育改革總諮議報告書」，建議的重點之一是，過去以人力需求的推估，管制高等教育科系的設立增減，應該「鬆綁」。
11	Education cannot be made merely to narrowly serve the economy.	教育不能「窄化」成只爲經濟服務，但現實的狀況是，
11	Yet in reality, "the reason most parents are willing to pay to put their children through university is certainly not that they hope they will become passionate seekers after truth, but to enable them to find good careers," says Providence University president Li Chia-tung bluntly.	「大部分家長之所以肯花錢讓孩子來念大學，絕不是希望孩子以後熱衷於真理的追求，而是爲了使孩子將來能找到好職業，」靜宜大學校長李家同明白地說。

Appendix B

More English-Chinese alignment results.

Sentence alignment based on length		
31	"The advocacy of core curriculum teaching is in itself a very important education for teachers." Lin Ku-fang says that when NHMC was set up it made broad-based education one of its founding principles, but discovered that attitudes were very hard to change, because "people today feel they are respected for their profession rather than their personality." Although when first studying an academic discipline one starts from a general outline, nonetheless one must be very well versed in a subject to teach it well.	「通識本身的提倡，對老師就是很重要的教育，」文化評論者林谷芳指出，南華成立時就把通識教育視爲創校理念，但還是發現觀念問題最難突破，因爲「現代人常覺得自己被尊重是因爲我的專業，而不是我的人。」
12	"There is a great sense of challenge about core curriculum teaching, but many people make the mistake of thinking it is very simple," says Lin.	雖然一門學問最初讀時是某某學導論，但真的得弄通，才教得好，「通識挑戰意味很濃，但大家都誤以爲很簡單。」
11	The scope of core curriculum teaching appears very broad, but it still has to start from the basics.	通識範圍看起來很廣，但還是由基礎出發，林谷芳認爲，任何學科都可以從「人與自然」、「人與人」、「人與超自然或自我」三個層次來看。
11	In Lin Ku-fang's view, any branch of academic learning can be viewed on the three levels of "man and nature," "man and man," and "man and the supernatural, or that which transcends self."	就以地球科學這門專業學科爲例，人所認識的自然是專業，提升到人與生態的互動、人與未來生命處境就是通識。
10	To take the example of earth science, a very specialized discipline, man's cognitive knowledge of nature is its specialist content, but to go a step higher and investigate the interactive relationship between man and ecology or man and the future condition of life requires a broad-based, multidisciplinary approach.	
Sentence alignment based on punctuation.		
21	"The advocacy of core curriculum teaching is in itself a very important education for teachers." Lin Ku-fang says that when NHMC was set up it made broad-based education one of its founding principles, but discovered that attitudes were very hard to change, because "people today feel they are respected for their profession rather than their personality."	「通識本身的提倡，對老師就是很重要的教育，」文化評論者林谷芳指出，南華成立時就把通識教育視爲創校理念，但還是發現觀念問題最難突破，因爲「現代人常覺得自己被尊重是因爲我的專業，而不是我的人。」
11	Although when first studying an academic	雖然一門學問最初讀時是某某學導

	discipline one starts from a general outline, nonetheless one must be very well versed in a subject to teach it well.	論，但真的得弄通，才教得好，
11	"There is a great sense of challenge about core curriculum teaching, but many people make the mistake of thinking it is very simple," says Lin.	「通識挑戰意味很濃，但大家都誤以爲很簡單。」
21	The scope of core curriculum teaching appears very broad, but it still has to start from the basics. In Lin Ku-fang's view, any branch of academic learning can be viewed on the three levels of "man and nature," "man and man," and "man and the supernatural, or that which transcends self."	通識範圍看起來很廣，但還是由基礎出發，林谷芳認爲，任何學科都可以從「人與自然」、「人與人」、「人與超自然或自我」三個層次來看。
11	To take the example of earth science, a very specialized discipline, man's cognitive knowledge of nature is its specialist content, but to go a step higher and investigate the interactive relationship between man and ecology or man and the future condition of life requires a broad-based, multidisciplinary approach.	就以地球科學這門專業學科爲例，人所認識的自然是專業，提升到人與生態的互動、人與未來生命處境就是通識。

Appendix C

Sentence alignment of English-Japanese parallel texts based on punctuation.

Sentence alignment based on punctuation		
Type	English text	Japanese Text
11	Liu Tseng-kuei, of Academia Sinica's Institute of History and Philology, once analyzed over 570 female names used during the Han dynasty in hopes it might shed some light on what the people of that time hoped to see in a woman.	中央研究院歴史語言研究所の副研究員である劉増貴さんは、漢代において女性に何が期待されていたかを理解するために、570 名余りの漢代の女性の名前を研究したことがある。
	, , .	、、。
12	It turns out that about two-thirds of the names examined were suitable for either women or men.	その結果、3分の2の名前が男でも女でも通用するものであることがわかった。漢代の女性の名前には実に力強いものも少なくない。
	.	、。。
21	Wang Mang, who usurped the throne in 9 AD, named his daughter Jie ("nimble and quick"). The daughter of the emperor Huan Di (132-167 AD) was named Jian ("solid and resolute") while her mother, the empress Deng, had the even more emphatic name of Mengnu, which means "fierce woman"!	王莽の娘の名は「倢」、後漢の桓帝の娘の名は「堅」といい、桓帝の時の皇后の名は、より直接的な「猛女」というものだったのである。
	, , (" ") . ( ) (" ") , , , " " !	「」、「」、、「」。
11	Says Liu, "These names show that society at that time had not yet come to hold the two sexes to such very different standards."	「この現象は、男性と女性の道徳行爲に対する社会の要求が、あまり違わなかったことを示しています」と劉増貴さんは言う。
	, " . "	「、、、」。
11	Although they were gradually beginning to use specifically feminine names alluding to a gentle and submissive nature, such traits as a resolute spirit and an agile, tough body were also seen as virtues in a woman.	当時、いわゆる女性的な名前もしだいに増えており、女性を低く見るという観念も確かにあったが、それでも女性が強くたくましくあることも肯定されていたのである。
	, , .	、、、、。
11	"The notion of the ideal woman being soft and weak was not so universally accepted then as it would later come to be."	「女性は弱くておとなしい方が良いとする考えは、後の時代のように絶対的なものではなかったようです」と劉増貴さんは言う。
	" . "	「、」。

References

Brown, P. F., J. C. Lai and R. L. Mercer, "Aligning sentences in parallel corpora," *Proceedings of the 29th conference on Association for Computational Linguistics*, Berkeley, CA, USA. 1991, pp. 169-176.

Chen, A. and and F. Gey , "Translation Term Weighting and Combining Translation Resources in Cross-Language Retrieval," *TREC 2001*.

Chen, K.H. and H.H. Chen, "A Part-of-Speech-Based Alignment Algorithm," *Proceedings of 15th International Conference on Computational Linguistics*, Kyoto, 1994, pp. 166-171.

Chen, S. F., "Aligning Sentences in Bilingual Corpora Using Lexical Information," *Proceedings of ACL-93*, Columbus OH, 1993, pp. 9-16.

Chuang, T., G.N. You and J.S. Chang , "Adaptive Bilingual Sentence Alignment," L*ecture Notes in Artificial Intelligence 2499*, pp. 21-30.

Collier, N., K. Ono and H. Hirakawa, "An Experiment in Hybrid Dictionary and Statistical Sentence Alignment," *COLING-ACL* 1998, pp. 268-274.

Danielsson, P. and K. Mühlenbock, Small but Efficient,"The Misconception of High-Frequency Words in Scandinavian Translation," *AMTA 2000*, pp. 158-168.

Déjean, H., É Gaussier and F. Sadat, "Bilingual Terminology Extraction: An Approach based on a Multilingual thesaurus Applicable to Comparable Corpora," *Proceedings of the 19th International Conference on Computational Linguistics COLING 2002*, Taipei, Taiwan, pp. 218-224.

Dolan, W. B., J. Pinkham and S. D. Richardson, MSR-MT, "The Microsoft Research Machine Translation System," *AMTA 2002*, pp. 237-239.

Gale, W. A. and K. W. Church , "A program for aligning sentences in bilingual corpora," *Computational Linguistics*, vol. 19, pp. 75-102.

Gey, F.C., A. Chen, M.K. Buckland and R. R. Larson , "Translingual vocabulary mappings for multilingual information access," *SIGIR 2002*, pp. 455-456.

Jutras, J-M., An Automatic Reviser, "The TransCheck System," *Proc. of Applied Natural Language Processing*, pp. 127-134.

Kay, M. and M. Röscheisen, "Text-Translation Alignment," *Computational Linguistics*, 19:1, pp. 121-142.

Ker, S.J. and J.S. Chang, "A class-based approach to word alignment," *Computational Linguistics*, 23:2, pp. 313-344.

Kueng, T.L. and K.Y. Su, "A Robust Cross-Domain Bilingual Sentence Alignment Model," *Proceedings of the 19th International Conference on Computational Linguistics*, 2002.

Melamed, I. D., "A portable algorithm for mapping bitext correspondence," *In The 35th Conference of the Association for Computational Linguistics (ACL 1997)*, Madrid, Spain,1997.

Melamed, I. D., "Bitext Maps and Alignment via Pattern Recognition," *Computational Linguistics*, 25(1), pp.107-130, March, 1999.

Moore, R.C., "Fast and Accurate Sentence Alignment of Bilingual Corpora," *AMTA* 2002, pp. 135-144.

Ribeiro, A., G. Dias, G. Lopes and J. Mexia," Cognates Alignment," In Bente Maegaard (ed.), *Proceedings of the Machine Translation Summit VIII (MT Summit VIII) – Machine Translation in the Information Age*, Santiago de Compostela, Spain, 2001, pp. 287–292.

Richards, J. et al., "Longman Dictionary of Applied Linguistics," Longman, 1985.

Simard, M., G. Foster and P. Isabelle, "Using cognates to align sentences in bilingual corpora," *Proceedings of TMI92,* Montreal, Canada, pp. 67-81.

Sinorama , The New University: Breaking down the Departmental Barriers, June, the 3rd article, 1998.

Wu, "Bilingual Collocation Extraction Based on Linguistic and Statistical Analyses," Master thesis, National Tsing Hua University, Taiwan, 2003.

Wu, D., "Aligning a parallel English-Chinese corpus statistically with lexical criteria," *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, New Mexico, USA, 1994, pp. 80-87.

Wu, J.C., K.C. Yeh, T.C. Chuang, W.C. Shei and J.S. Chang, "TotalRecall: A Bilingual Concordance for Computer Assisted Translation and Language Learning," *ACL2003 workshop.*

Yang, Y., "Researches on Punctuation Marks," Tien-Chien Publishing, Hong Kong.

Yang, C. and K. Li, "Automatic Construction of English/Chinese Parallel Corpora," *Journal of American Society of Information Science and Technology*, 54(8), 2003, pp 730-742.