Alignment of English-Chinese Parallel Corpora and its Use in Cross-Language Information Retrieval

Aitao Chen*, Fredric Gey†, and Hailing Jiang*
*School of Information Management and Systems
†UC Data Archive & Technical Assistance (UC DATA)
University of California at Berkeley, CA 94720, USA
{aitao, hjiang1}@sims.berkeley.edu, gey@ucdata.berkeley.edu

Abstract

Bilingual lexicons are valuable resources for machine translation, multi-lingual information retrieval, human interpreters, translators, and foreign language learners. This paper presents a method to align two Chinese/English parallel corpora from the document level down to the sentence level. One word and one phrase bilingual lexicons are derived from the aligned sentences. The usability of the bilingual lexicons are tested in English-to-Chinese cross-language information retrieval.

Keywords: parallel text alignment; bilingual lexicon construction; cross-language IR.

1 Introduction

The alignment of parallel texts requires that for every fragment of text in the source language the corresponding text in the other language be located. Alignment can be done at various levels such as the document, paragraph, sentence, and word. The subject of aligning parallel text at the sentence and word level has received a lot of attention during the past decade. There is a large literature on aligning parallel texts [16, 12, 9, 11, 2, 19, 20, 7, 10, 14, 15, 17], The techniques on parallel text alignment include length-based [9], lexicon-based [11], combination of length and lexicon knowledge [2, 20, 10], and others [7]. Chapter 1 in [17] presents a survey of parallel text processing. Interested readers can also find a recent summary of various parallel text alignment methods in chapter 13 in [13], including Gale and Church's [9] method that we adapted for Chinese-English parallel text alignment. The alignment of parallel text at the word level results in a bilingual lexicon which is useful in such areas as machine translation, multi-lingual information retrieval, bilingual lexicography, and so on. A bilingual lexicon is also a valuable resource for human interpreters, translators and people who are learning foreign languages.

We obtained two Chinese-English parallel corpora. Our main focus here is to derive a Chinese-English bilingual lexicon from the parallel corpora through parallel text alignment and then use the resultant lexicon in English-Chinese cross-language information retrieval. We are particularly interested in deriving translations of phrases since the phrase translations are often missing in bilingual dictionaries and some phrase translations are difficult or simply cannot be obtained from the word-by-word translations. For example, "human rights" is expressed as "人权" in Chinese; "White Paper" is "白皮书" in Chinese; and "Hong Kong dollar" is "港元" in Chinese.

The next four sections discuss the alignment of two Chinese/English parallel corpora at the document, paragraph, sentence, and word level, respectively. Then the derived bilingual lexicon is applied to English-to-Chinese cross-language retrieval. Our conclusions are then summarized.

2 Alignment at document level

We downloaded 59,585 news articles from the news archive of the daily Press Release of the Hong Kong Special Administrative Region Government. The news articles cover the period from April, 1998 through March, 2001. Most of the news articles are published in both Chinese and English. We will call this collection of news articles the *Hong Kong News Corpus* from now on. Each file contains one single news article in Chinese or in English. However, the names of the files containing the same news article in English and in Chinese are not related so one cannot tell from the file names alone which file contains the Chinese news article and which file contains the same new article in English. Our first task is to sort out which pair of files containing the same news article. We initially developed a simple algorithm to match documents based on the observation that the date, time, and numeric expressions in Chinese often have rigid translations in English. For example, the date 十月二十七日 is translated into October 27, or Oct 27; 星期日 is translated into Sunday; and the numeric expression 一千七百九十九 is expressed in 1,799 in English. Also the proper nouns, such as the personal names, organization names, and place names have rigid translations. Initially we used date, time, and numeric expressions to index the documents. To decide the file that most probably contains the English translation of a Chinese news article, we first translated the Chinese date, time, and numeric expressions into English. Then we computed the similarity between the Chinese document and all English news articles released in the same day. The content of an English article is represented by the English date,

time, and numeric expressions found in the English article. And the content of an English translation of a Chinese document is also represented by date, time, and numeric expressions found in the translation. Since both Chinese and English versions of news releases were published on the same day, we organized the files by date, then aligned the news articles for a single day each time. We first indexed all the English news articles published on each day Then we indexed the English translations of all the Chinese news articles published on the same day as the English articles. We computed the similarities between every pair of English news article and English translation. A Chinese document is paired with the English news article which has the largest similarity with the English translation of the Chinese document. The similarity between a pair of documents is computed using the document ranking formula presented in [3]. This method works well when a document has several in date, time, and numeric expressions, but poorly when a news article contain few or no date, time, and numeric expressions. One enhancement is to include proper names in indexing since the proper names also have fixed translations. Of course, to apply this technique would require the identification of proper names in both English and Chinese, and the availability of a bilingual dictionary containing the proper names and their translations.

To evaluate the effectiveness of this simple method, we aligned all the news articles published in March, 2001 using the date, time, and other numeric expressions in the documents. The alignment produced 808 document pairs from 1,673 documents. Note that some of the news articles are published in Chinese only or English only. We randomly selected 202 document pairs for evaluation and found only 63 pairs were correctly matched.

In our second attempt to align documents, we used the LDC English/Chinese bilingual dictionary to translate the Chinese news articles published on each day into English. Firstly, we separately indexed the Chinese documents and the English documents for the same day. The index terms are created using the dictionary-based longest matching method. Each index term in a Chinese document is looked up in the LDC bilingual dictionary, a Chinese-to-English wordlist (version 2.0) compiled by the Linguistic Data Consortium. We obtained the bilingual wordlist from the web site at http://morph.ldc.upenn.edu/Projects/Chinese/. The wordlist consists of a list of Chinese words, paired with a set of English words. The wordlist has some 128,000 entries. When the dictionary includes multiple translations for a single Chinese term, we selected the English translation that occurs most frequently in the English news articles for the same day. We then take the English translation of a Chinese news article as query, and compute the similarity value between the Chinese news article and all the English news article published on the same day. The similarity between two documents was estimated using the similarity function presented in [3]. The English article of the largest similarity value is paired with the Chinese news article.

We used the second method to align all the articles. The document alignment resulted in 25,199 pairs of English/Chinese documents. We randomly selected 500 pairs of aligned documents, and manually checked the correctness of the alignment. We found 472 out of the 500 pairs are correct. The accuracy or precision is about 94.4%.

The second Chinese-English parallel corpus is created by Foreign Broadcasting Information Service. We will call this collection of documents the *FBIS corpus*. This corpus is about 60MB in size, consisting of 5,992 files. However, we found 2,976 of the files were duplicate documents. So there are only 3,016 unique documents in the corpus. Among the unique documents, 52 of them are either incomplete or corrupted in encoding. The actual number of documents used in alignments is 2,964. Most of the documents are news articles. Each file contains one Chinese document and its English translation. About half of the documents are encoded in UTF-8, and the other half either in GB or Big5 coding scheme. We converted the files in UTF-8 and GB into Big5.

The average English to Chinese length ratio at the document level is 2.30. While the English translations are well formatted, the original Chinese texts are not. Some Chinese texts are placed in the middle of a page. In some Chinese text, the Chinese characters are separated by a blank space. We ignored the blank space character and the new line character appearing in the Chinese text when we computed the length of a Chinese document. Since a Chinese document and its English translation are stored in the same file, there is no need to align documents.

3 Alignment at paragraph level

Since some of the documents in the FBIS corpus are rather long, we did not attempt to align the whole document at the sentence level in a single step because the chances of having incorrect alignments at the sentence level would be greater. Instead we proceeded in two steps. We first aligned the documents at the paragraph level. Then we aligned the paragraphs at the sentence level.

We adapted Gale and Church's alignment algorithm which is length-based [9]. The algorithm is based on the observation that long sentences in one language tend to be translated into long sentences in the target language; and short sentences tend to be translated into short sentences in the target language. So the lengths of the source text and its translation are highly correlated. For example, Gale and Church reported that the length ratio of English text and its French translation is approximately one. In their model, length of a sentence is measured in term of characters instead of words because length difference varies less widely.

The model works well on language pairs where the length of the source text and the length of its translation are more or less the same, such as the English-French language pair. In adapting Gale and Church's method to align Chinese/English documents at the paragraph level, we included two changes. Firstly, we measure sentence length in terms of bytes for both English and Chinese texts rather than in term of characters as in the original model. Secondly, we extended the length of the Chinese text to about the same as that of the English text by computing the length ratio of the English document over the Chinese document in the same pair, and then multiplying the length of each Chinese paragraph by the document length ratio so that both the English and Chinese documents have roughly the same length. On the average, the English translation of the Chinese text is about 2.3 times as long as the source Chinese text in the FBIS corpus. Without the length adjustment for the Chinese text, the precision of paragraph alignment would not be very good since the alignment algorithm assumes that the source text length and its translation text length are more or less the same which is clearly not true for Chinese/English language pair.

Before we aligned the documents at the paragraph level, we first identified the paragraphs in each document, which is a relatively easy task. The documents in the FBIS Chinese/English corpus are marked with XML tags. The English translation of the Chinese text is explicitly marked, and the translation of the Chinese title is also marked. However, the title of a Chinese article, name(s) of the reporter(s), and their affiliations are not separated from the main body of the Chinese news articles. We wrote some utility programs to take the title, the reporter names, and their affiliations out of the main body of the news articles. For those documents in which it is difficulty to accurately identify the title, reporter names and their affiliation, we manually took these text out of the body of the Chinese text. Data cleaning is necessary since Gale and Church's algorithm aligns parallel texts solely based on their lengths.

The alignment of the 2,964 documents in the FBIS corpus at the paragraph level generated 21,940 paragraph pairs. A sample of 500 paragraph pairs were randomly selected for evaluation. 456 (91.2%) paragraph aplignments are correct, 19 (3.8%) are partially correct, and 25 (5%) are incorrect. A paragraph alignment is considered correct when the English text in an aligned paragraph pair is the complete translation of the Chinese text in the same pair. An alignment is considered partially correct if the English text in an aligned paragraph pair has at least one sentence that is the correct translation of some part of the Chinese text in the same pair. When no sentence in the English text of an aligned paragraph pair is the correct translation of any part of the Chinese text, the paragraph alignment is considered incorrect.

Compared with the documents in the FBIS Chinese/English corpus, the documents in the Hong Kong News corpus are much smaller on the average. So we did not align the documents at the paragraph level. We aligned the documents in the Hong Kong News corpus at the sentence level in one step, treating each document as one paragraph.

4 Alignment at sentence level

entenceAlignment For the Hong Kong News corpus, we used Gale and Church's method to align documents directly at the sentence level since most of the documents are small. The changes to the original algorithm included the measurement of length in byte and Chinese sentence length adjustment as was done in aligning the documents in the FBIS corpus at the paragraph level.

Some of the sentences in the FBIS corpus are rather long, and in that case, the long Chinese sentences are often translated into two or more English sentences. We even noticed a case in which one Chinese sentence is translated into six English sentences. It is much more common to see a Chinese sentence being translated into two or three English sentences than the case where two or three Chinese sentences are translated into one English sentence. Gale and Church's method considers the following six cases only: insertion, deletion, replacement, expansion, contraction, and two-to-two correspondence. Only one-to-two expansion and two-to-one contraction are considered in Gale and Church's method. Their method is not able to deal with the cases in which one sentence is translated into three or more sentences, and the cases in which three or more sentences together are translated into a single sentence. Since the contraction case rarely happen in the FBIS Chinese/English bilingual corpus, the ignorance of the cases where three or more Chinese sentences together are translated into one English sentence need not concern us here. However, it is not unusual to see cases where one Chinese sentence in the FBIS corpus is translated into three or more English sentences.

From the aligned documents in the Hong Kong News corpus, we generated 277,592 pairs of aligned sentences. We randomly selected 309 pairs of aligned sentences from the Hong Kong News Corpus and manually checked for their correctness. We found 235 cases (which is only 76.05%) were correct, 24 cases were partially correct, and 50 cases were incorrect.

From the aligned paragraphs in the FBIS corpus, a total of 66,621 pairs of aligned sentences were generated using the modified method. To compare the alignment accuracy at the sentence level using the original Gale and Church's method and the modified method with Chinese length adjustment, we also aligned the paragraphs using the original method. We randomly selected 500 aligned sentence pairs from all the sentence pairs generated using the original method, and another 500 aligned sentence pairs from all the sentence pairs generated using the modified method. We manually checked for the alignment correctness. The evaluation results are presented in ta-

ble 1. A sentence alignment is considered partially

Method	Correct Cases	Partially Correct Cases	Incorrect Cases
Original	290 (58.0%)	99 (19.8%)	111 (22.2%)
Modified	391 (78.2%)	50 (10%)	59 (11.8%)

Table 1: Accuracy of sentence alignment for the original and modified methods. The sentences were aligned from imperfectly aligned paragraphs. Sample size is 500 in both cases.

correct when the English text contains at least one sentence that is a correct translation of some part of the Chinese text in the aligned sentence pair. The sentences were aligned from the imperfectly aligned paragraphs. The accuracy would be higher if the paragraphs from which the aligned sentence pairs were generated were perfectly aligned. As mentioned in section 3, the precision of the paragraph alignment is 91.2%. Of the 50 partial matchings in our second sample of 500 sentence pairs, we found 12 cases where one Chinese sentence was translated into three English sentences; one case where one Chinese sentence was translated into four English sentences; and two cases wehre one Chinese sentence was translated into six sentences. We found the subtitles and the text (e.g. the name of news agency, date, and name of reporter) preceding the leading sentence often result in partial or incorrect alignments.

5 Construction of translation lexicons

Our goals in aligning the Chinese/English corpora is two-fold: Firstly, we would like to derive a bilingual Chinese/English lexicon. We are especially interested in deriving phrase translations because often they are absent in bilingual dictionary, and some phrase translations are difficult to obtain from the translations of the words in the phrases. Phrasal translation is also appealing to cross-language information retrieval since the translation of a phrase as a whole is usually unambiguous, thus eliminating the problem of disambiguation after translation as often happens in single word translations. Secondly, we want to use the resultant bilingual lexicon as a translation aid in performing Chinese/English cross-language information retrieval.

In this section, we will first describe our method of recognizing noun phrases in English documents. The recognition of noun phrases in Chinese documents depends upon their existence in the Chinese segmentation dictionary. We did not attempt to identify noun phrases in Chinese documents by other means, such as tagging and parsing the Chinese text. We will then describe the method to find the most likely Chinese translations of an English word or phrase using the Chinese/English corpora aligned at the sentence level. Our interest here is not to align words between each sentence pair in the sense as used in statistical machine translation. We are trying is to determine the most probably Chinese translation(s) of an English word or phrase considered in isolation from the rest of the sentence.

We attempted to identify only the simple noun phrases in the English documents from the FBIS and Hong Kong News corpora. By simple noun phrases we mean the ones that end with a noun preceded by any number of noun or adjective pre-modifiers and without any post-modifiers. Simple noun phrases were extracted from the English documents in two steps. Firstly, the English documents are tagged with part of speech using Brill's rule-based part-of-speech tagger [1]. Secondly, simple noun phrases are extracted based on the pattern, or sequence, of the part-of-speech tags assigned to words. We have used a simple three-state automaton where the edges are labeled with part-of-speech tags. A sample of tagged sentence is shown below.

> With/IN the/DT implementation/NN of/IN these/DT and/CC other/JJ measures/NNS ,/, air/NN pollution/NN by/IN vehicle/NN emission/NN s/PRP will/MD be/VB significantly/RB alleviated/VBN in/IN the/DT coming/VBG years/NNS ./.

Each word is followed by its part-of-speech tag. The tags NN and NNS represent singular nouns and plural nouns, respectively; NNP represents the proper name, and JJ represents adjective. Then the tagged text is passed to a noun phrase recognizer represented in a three-state automaton for noun phrase extraction. The recognizer detects simple noun phrases based on the pattern of the tags. The noun phrase patterns we used to extract noun phrases can be concisely specified in a three-state automaton as shown in Figure 1. The initial state is 0 and the final state is 2. Any words tagged with part-of-speech tags NN, NNS, NNP, NP and NPS are represented by the label NOUN in the graph, and words tagged with JJ, JJR, and JJS, which are the positive, comparative and superlative form of an adjective, are represented by the label ADJ. Any sequence of words whose part-of-speech tags completes the path from the initial state to the final state will be extracted as a noun phrase, excluding the single-word nouns. The noun phrases extracted from the above tagged sentence include *other measures*, *air pollution*, and *vehicle emission*.

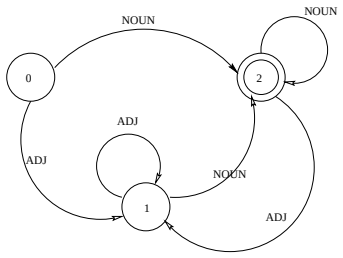

Figure 1: Simple noun phrase automaton

The Chinese text in the bilingual corpora were segmented using the longest matching method with a segmentation dictionary of approximately 200,000 entries, including noun phrases and proper names. A Chinese noun phrase listed in the segmentation dictionary has a good chance of being recognized during segmentation since we used the longest matching method. The noun phrases not included in the segmentation dictionary will not be identified in the Chinese text. One could apply the same method used for noun phrase identification as described in this section to extract noun phrase in the Chinese text if a Chinese part-of-speech tagger is available.

Each noun phrase in the English documents is treated as a single term in word alignment. The intuitive idea on finding the most probable translation of a source term is based on the co-occurrence pattern in the aligned sentence pairs. When a pair of source and target terms (words or phrases) often co-occur in the aligned sentence pairs, and it rarely happens that when the source term occurs in a sentence pair where the target term is not found, then there is a good chance that the target term might be the equivalent of the source term. One approach to finding the most probable translation equivalent in the target language is to compute the association degree between the source term and all the terms in the target language that co-occur with the source term at least once, then choose the target terms with the highest association values as the translation equivalents of the source term. To compute the association degree between a pair of words, one in the source language and the other in the target language, one could create a two-by-two contingency table showing the number of cases (i.e., aligned sentence pairs) where both terms occur together, the number of cases either term occurs alone, and the number of cases neither term occurs. The association degree could be measured using such statistic as Chi-square and mutual information [8]. Here we used an association measure called likelihood ratio test statistic developed by Dunning [4] to compute the association degree between a pair of Chinese/English words.

If the English word 'E' is a translation of the Chinese word 'C', then one would expect that when the Chinese word 'C' is present in a Chinese sentence, its English translation 'E' would also appear in the paired English sentence. From the aligned sentences,

	Chinese word	
English word	a	b
	c	d

Table 2: A contingency table for a pair of words.

we created a contingency table for every pair of Chinese/English words as shown in table 2, where a is the number of aligned sentences containing the pair of Chinese/English words; b is the number of aligned sentences containing the English word, but not the Chinese word; c is the number of aligned sentences containing the Chinese word, but not the English word; and d is the number of aligned sentences containing none of the word pair. The association score between a Chinese word 'C' and an English word 'E' was computed as follows [4]

$$\begin{aligned} W(C,E) &= 2[logL(p_1, a, a+b) + logL(p_2, c, c+d) \\ &\quad -logL(p, a, a+b) - logL(p, c, c+d)] \end{aligned}$$

where $logL(p, n, k) = klog(p) + (n-k)log(1-p)$, $p_1 = \frac{a}{a+b}$, $p_2 = \frac{c}{c+d}$, and $p = \frac{a+c}{a+b+c+d}$. The number of aligned sentence pairs generated from both parallel corpora is about 313,890. The Chinese vocabulary size is 69,678 words and the English vocabulary size is 48,176 words. Table 3 presents 40 English noun phrases or proper names automatically extracted using our noun phrase recognizer that occur most frequently in the corpora. The Chinese translations of these English phrases are shown in column 3 in table 3. Some of the Chinese translations are not complete because the Chinese noun phrases are missing in our segmentation dictionary. The Chinese phrases

	English Phrases	Chinese Translations
1	hong kong	香港
2	united states	美国
3	taiwan independence	台独
4	chief executive	行政长官
5	standing committee	常委会
6	physical exertion	体力
7	last year	去年
8	human rights	人权
9	chinese people	中国人民
10	heavy traffic	繁忙
11	one-china principle	一个中国原则
12	chinese government	中国政府
13	western region	西部
14	political work	政治
15	basic law	基本法
16	state council	国务院

Table 3: The top 16 most frequent English phrases and the Chinese translations automatically extracted from the Chinese/English corpora.

corresponding to the English phrases *physical exertion*, *heavy traffic western region*, and *political work* are not complete because the complete phrases are missing in our segmentation dictionary. The other 12 Chinese phrases presented in table 3 are complete and are correct translations of the corresponding English phrases.

6 Related work

Wu [20] also applied Gale and Church's method to a parallel English-Chinese corpus consisting of the parliamentary proceedings of the Legislative Council of Hong Kong. He made two similar modifications to the original Gale and Church's method: 1) the length of the Chinese text was measured in byte, and 2) the English text length was reduced by a factor of .506, which is the average number of Chinese characters generated by each English character [20]. Despite much greater variance in length between English and Chinese, Wu found the modified method worked well. When lexicon cues were incorporated, the performance on English-Chinese alignment was close to that on English-French alignment as reported by Gale and Church [9]. Our results are not as good as those in [20]. We believe there are a number of reasons: 1) we did not use any lexicon in alignment; 2) our corpus, especially the FBIS corpus, has different characteristics; 3) the sentence alignment for the FBIS corpus was done on automatically, but not perfectly, aligned paragraphs. On the average, the length ratio of a Chinese document over its English translation in the FBIS corpus is 2.30. In our randomly selected 500 aligned sentence pairs, we found 12 cases where a Chinese sentence was translated into three English sentences; 1 case where a Chinese sentence was translated into four English sentences; and even 2 cases where a Chinese sentence was translated into six English sentences. For the same number of test cases, there was only one case in [20] where one Chinese sentence was translated into three English sentences.

Fung presented a method in [5] for deriving bilingual lexicon from non-parallel English-Chinese corpus using the heterogeneity contexts of words in the corpus, and a method in [6] based on word frequency and position information for compiling noun and proper noun translations from noisy, but unaligned, parallel texts.

7 Cross-Language Retrieval

We tested the usability of the lexicon derived from the two Chinese/English corpora as described above in English-to-Chinese cross-language information retrieval using the test document/query set for the Cross-Language Information Retrieval Track in TREC-9 [18]. The test collection consists of 25 new topics and 127,938 documents from three newspapers, namely the Hong Kong Commercial Daily, Hong Kong Daily News, and Takungpao. The topics are written in English with Chinese translations, and contain *title*, *description*, and *narrative* fields.

We created two bilingual lexicons from the Chinese/English corpora, one consisting of words only and the other consisting of words and phrases. Before query translation, the English topics were tagged and simple noun phrases were extracted as described above. The noun phrases and the single-word terms found in the original topics were translated into Chinese by dictionary lookup in the derived lexicons. The English phrases were looked up in the phrase lexicon. And when a phrase was not found in the phrase lexicon, it was translated in word-by-word by looking up the constituent words in the word lexicon. Other single-word terms were also looked up in the word lexicon.

	Monolingual	CLIR (LDC)	CLIR (Derived)
average precision	0.2936	0.1680	0.1855
% of mono		57.22%	63.18%

Table 4: Evaluation results for one Chinese monolingual run and two English to Chinese cross-language retrieval runs.

The TREC-9 Chinese documents were segmented into Chinese using the longest matching method with the same segmentation dictionary as mentioned above. The Chinese translations of the original English topics were searched against the Chinese document collection. And the top-ranked 1000 documents were used to compute recall-precision. The experimental results are presented in table 4. Column 2 in table 4 shows the evaluation results for the monolingual Chinese retrieval where both the documents and queries were broken into overlapping bigrams in indexing. Column 3 shows the English-to-Chinese cross-language retrieval results using the LDC bilingual lexicon; and the last column in table 4 presents the evaluation result for the English-to-Chinese cross-language retrieval using our automatically derived lexicons from the Chinese/English parallel corpora.

8 Conclusion

This paper has presented an interesting case where information retrieval techniques are applied in parallel text alignment and parallel text alignment is then applied in cross-language information retrieval. We have presented a method to automatically construct bilingual word and phrase lexicons from parallel corpora through aligning parallel text down to the sentence level. We have introduced two modifications to the sentence alignment algorithm developed by Gale and Church in adapting their method to aligning Chinese/English parallel text. The modifications have significantly improved the precision of sentence alignment.

The automatically derived word and phrase lexicons are used to translate English queries into Chinese in the English-to-Chinese cross-language retrieval evaluation. The cross-language retrieval performance shows that the derived lexicons are as useful as the manually constructed bilingual lexicon in cross-language retrieval.

9 Acknowledgements

This research was in part supported by DARPA under research grant N66001-00-1-8911 as part of the DARPA Translingual Information Detection, Extraction, and Summarization Program (TIDES). We would like to thank the anonymous referees for constructive comments.

References

[1] Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.

[2] Stanley F. Chen. Aligning sentences in biligual corpora using lexical information. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 9–6, 1993.

[3] W. S. Cooper, A. Chen, and F. C. Gey. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 57–66, March 1994.

[4] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19:61–74, March 1993.

[5] Pascale Fung. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *The 3rd Annual Workshop on Very Large Corpora*, pages 173–183, Boston, Massachusette, 1995.

[6] Pascale Fung. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics*, Boston, Massachusette, 1995.

[7] Pascale Fung and Kenneth Church. K-vec: A new approach for aligning parallel texts. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 1096–1102, Kyoto, 1994.

[8] William A. Gale and Kenneth W. Church. Identifying word correspondences in parallel texts. In *Proceedings of the Fourth DARPA Speech and Natual Language Work shop*, pages 152–157, Pacific Grove, CA, 1991.

[9] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19:75–102, March 1993.

[10] Masahiko Haruno and Takefumi Yamazaki. High-performance bilingual text alignment using statistical and dictionary information. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 131–138, Santa Cruz, California, 1996.

[11] Martin Kay and Martin Roscheisen. Text-translation alignment. *Computational linguistics*, 19:121–142, March 1993.

[12] G. F. Foster M. Simard and P. Isabelle. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81, Montreal, Canada, June, 1992.

[13] Christopher D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[14] I. Dan Melamed. Models of translational equivalence among words. *Computational linguistics*, 26:221–249, June 2000.

[15] Franz J. Och and Hermann Ney. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrucken, Germany, July 2000.

[16] J. C. Lain P. F. Brown and R. L. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176, Berkeley, CA, 1991.

[17] Jean Veronis, editor. *Parallel text processing : alignment and use of translation corpora*. Kluwer Academic Publishers, Boston, 2000.

[18] E. Voorhees and D. Harman, editors. *The Ninth Text Retrieval Conference (TREC-9)*, National Institute of Standards and Technology, Gaithersburg, MD, 2000.

[19] Dekai Wu. Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 80–87, Las Cruces, New Mexico, 1994.

[20] Dekai Wu and Xuanyin Xia. Learning an english-chinese lexicon from a parallel corpus. In *Proceedings of the first Conference of the Associatioin for Machine Translation in the Americas*, pages 206–213, Columbia, Maryland, July 1994.