Automatic Acquisition of a High-Precision Translation Lexicon from Parallel Chinese-English Corpora

Zhao-Ming Gao*
Academia Sinica

This paper presents a hybrid approach to deriving a translation lexicon from unaligned parallel Chinese-English corpora. Based on the observation that the English translation of a Chinese compound often involves two or more English words, a heuristic is presented that accepts Chinese-English word pairs as correct if two consecutive English words are proposed to correspond to the same Chinese word in a statistical or dictionary-based approach. In addition, document-external word distributions are utilised to measure word pair co-occurrences in the whole corpus. Words with low co-occurrence ratios are then filtered out. The two proposed methods can derive a translation lexicon of more than 94% precision.

1. Introduction

The compilation of a large translation lexicon is very costly, laborious, and time-consuming. Automation of all or most of the compilation process is therefore highly desirable. This goal has been made possible by the availability of large parallel corpora, machine-readable bilingual dictionaries, and statistical algorithms that have been proposed in recent years. In this paper, we address the problems encountered by statistics-based and dictionary-based approaches in deriving and augmenting a translation lexicon. We advocate a hybrid approach combining statistical and dictionary information to increase the recall. In addition, we suggest using proximity and global distributions of word pairs to enhance the precision.

2. The K-vec Algorithm

Fung and Church (1994) propose a simple algorithm to find word correspondences from unaligned parallel texts. The basic idea is that a true word pair should have similar distributions in terms of the position of its occurrence in the text. To estimate the similarity of co-occurrence, the parallel texts are split into the same number of segments (K) and the distributions of each word are represented in a 1...K binary vector. For instance, suppose the Chinese and English texts are divided into ten segments. Suppose further that the Chinese word 大學 *daxue* occurs ten times, with the first 3 occurrences in the fourth segment and the remaining 7 occurrences in the seventh segment and that the English word *university* appears twelve times, with the first 4 occurrences in the fourth segment and the remaining 8 occurrences in the seventh segment. Using the K binary vectors, the distributions of both the Chinese and English words in question can be represented as <0,0,0,1,0,0,1,0,0,0>. Mutual information (MI) and t-score are then used to estimate the correlation of a proposed word correspondence. The mutual information and t-score of the word pair are defined in (1) and (2).

(1)

$$MI(V_c, V_e) = \log_2 \frac{P(V_c, V_e)}{P(V_c)P(V_e)}$$

$$P(V_c) = \frac{a+b}{K}$$

$$P(V_e) = \frac{a+c}{K}$$

$$P(V_c, V_e) = \frac{a}{K}$$

* Chinese Knowledge Information Processing Group, Institute of Information Science, Academia Sinica, Nankang, Taipei 115, Taiwan. E-mail: imgao@hp.iis.sinica.edu.tw

where a is the number of pieces of segments in which both the Chinese and the English words occur; b is the number of pieces of segment where only the Chinese word is found; c is the number of pieces of segment where only the English word is found.

$$(2).\quad t(V_c, V_e) = \frac{P(V_c, V_e) - P(V_c)P(V_e)}{\sqrt{\frac{P(V_c, V_e)}{K}}}$$

The t-score is introduced to filter out word pairs with low frequency which happen to co-occur in the same segment by chance. The threshold value of MI and t-score are set to be 0 and 1.65, respectively. Only word pairs which are higher than the predetermined threshold values and in the frequency range 3-10 are considered to be potential mutual translations. Fung and Church (1994) observe that K, the number of segments in the bilingual texts, has a decisive factor in the performance of the algorithm. If K is too large, there will be many spurious word pairs. On the other hand, if K is too small, very few word pairs can be identified. They thus suggest that K be set to the square root of the length of the bilingual texts.

We reimplemented the K-vec algorithm and tested it with seven articles of the *Sinorama* Chinese-English parallel corpus, which consists of bilingual articles from the *Sinorama* Magazine. Since there are no delimiters between words in Chinese, the Chinese texts were first preprocessed by the word segmentation program reported in Chen and Liu (1992).

Table 1 summarises the results of the experiments. It indicates that the number of word pairs proposed by K-vec depends on the length as well as the nature of the text. Although K-vec invariably performs poorly with small texts, there are significant discrepancies between bilingual texts of similar length. Take texts 5 and 6 for example, which are of similar length but for which the results differ radically. Equally noticeable in Table 1 is the fact that longer texts do not necessarily have better performance in terms of recall or precision, as can be seen from comparisons between texts 2 and 5, as well as 6 and 7.

Table 1. The Influence of Text Length on the Performance of K-vec

No	Chinese length	English length	proposed pairs	correct pairs	precision
1	990	1221	0	0	0.00
2	1582	1754	6	3	0.50
3	2305	2743	2	2	1.00
4	3859	4805	8	6	0.75
5	4428	5138	5	2	0.40
6	4605	5557	40	24	0.60
7	5155	6414	24	12	0.50

Since none of the bilingual articles chosen involved additions or deletions of a large segment, we suspect that the huge disparity of performance manifested in Table 1 may be attributed to differences in text structures. Quantitatively speaking, if a bilingual text is full of recurrent terms or proper names, the co-occurrence ratio of a true word pair will be high, as they do not have morphological variants, synonyms, or hyponyms which discount the similarity measures. As a result, K-vec can perform better. If, however, the lexical patterning of a bilingual text involves relatively few identical repetitions, the performance of K-vec will inevitably be poor. In other words, the cohesive devices employed in a bilingual text greatly influence the performance of K-vec.

K-vec is a typical example which shows the strength and limitations of statistical word alignment algorithms. In fact, all the statistical algorithms that have been proposed to extract word correspondences from unaligned parallel corpora including Kay and Röscheisen (1993), K-vec, and DK-vec (Fung and McKeown (1994, 1997)) cannot be exempt from the limitations on text length and frequency (c.f. Haruno and Yamazaki's (1996) critique of Kay and Röscheisen (1993), Jones and Somers' (1995) experiments on K-vec, and Somers and Ward's (1996) evaluation of DK-vec based on several parallel corpora). Due to the shortcomings of statistical approaches, several researchers (e.g. Kumano and Hirakawa (1994), Utsuro et al. (1994), Haruno and Yamazaki (1996)) have advocated combining statistical and linguistic information. In the following section, we explore a dictionary-based approach and show why it is necessary to integrate linguistic with statistical information.

3. Why is it Non-trivial to Use Bilingual Dictionary Lookup for Word Alignment?

Contrary to one's intuition, using bilingual dictionary lookup to find word correspondences is non-trivial. We conducted a small experiment using English-to-Chinese dictionary lookup on the basis of exact string matching. The result showed that of 212 English words our bilingual dictionary actually only found 17 translations, five of which were contextually incorrect. In other words,.the precision of exact matching was only 70.59%, while the recall was as low as 5.66%.

The low recall was due to three factors. First, the bilingual dictionary we used was not comprehensive. Second, we did not use morphological processing and simply removed the most productive suffixes such as *-ed*, *-ing*, *-er*, and *-est*. Third, words might not be the smallest units for translation. Often, translations are done chunk by chunk, where a chunk might consist of a phrase, collocation, fixed expression, or even a sentence. Furthermore, some constructions or patterns must be regarded as translation templates that cannot be decomposed into smaller units.

Since the recall of dictionary lookup based on exact matching was too low, we experimented on using inexact (i.e. partial) matching. In the same sample paragraph of 212 English words, English-to-Chinese dictionary lookup based on partial matching suggested 172 word correspondences, 51 of which were correct, obtaining 29.65% precision and 24.05% recall. This result showed that inexact matching has higher recall (24.05% vs. 5.66%) but lower precision (29.65% vs. 70.59%) than exact matching. The choice of exact or inexact matching is thus a trade-off between precision and recall. The following question naturally arises. How can we filter out spurious word pairs if inexact matching is adopted?

4. Using Positional Information to Filter Out Unlikely Word Correspondences

Positional information plays an important role in distinguishing which word correspondence is more likely. The notion of applying positional difference information to word alignment has been used by several researchers in various forms, e.g. Dagan et al. (1993), Fung and McKeown (1994, 1997), Jones and Somers (1995). Intuitively, we would expect that the positional ratio of a word in the source text to its translation in the target text should not differ too much if there are not many omissions or additions in the translation process. This intuition can be expressed in (3).

(3). $$\left| \frac{x_i}{n_e} - \frac{y_i}{n_c} \right| \leq k$$

where x_i and y_i are the offsets from the beginning of the English and Chinese texts in words and n_e and n_c are the total number of words in the English and Chinese texts.

The value of k is in inverse proportion to the length of the text and must be empirically determined. In our experiment with texts of 3000 - 4000 words, *k* was set to 0.02. Using the constraint in (3), we observed that about 95% of correct word correspondences were retained, while approximately 60% incorrect word correspondences were excluded. Positional difference information significantly reduced a lot of uncertainties. However, there are still a lot of spurious word correspondences that haven't been filtered out. Obviously, we need another method to clean the translation lexicon.

5. Using Proximity to Find the Translations of Chinese Compounds

It has been noticed that there is a large number of Chinese compounds whose English translations involve more than one word. This characteristic can be easily utilised to identify translations of Chinese compounds from the output of K-vec and dictionary lookup based on partial matching. Since K-vec extracts bilingual word pairs which co-occur in the same segment more often than by chance, if it associates the same Chinese word with two or more English words which are adjacent to each other, then we can reasonably infer that these English words are translations of the Chinese word. Given that tables of word position·indexes of the bilingual text are required by K-vec, extraction of the translations of Chinese compounds is very straightforward. The second and third columns (i.e. the Chinese and English words) of Table 4 are the word pairs extracted by K-vec. Using these word pairs and word position indexes, we can extract the translations of Chinese compounds such as (口試 *koshi* ⇔ *oral exam*) and (性騷擾 *xingshaorao* ⇔ *sexual harassment*). Similar approaches have also been employed in Fung and McKeown (1994, 1997).

The same principle can be easily applied to the output of dictionary lookup based on partial matching. If the translations of two or three consecutive English words partially match different characters of the same Chinese word in the text then they are very likely to be correct word pairs, as the probability of the translations of two consecutive English words coincidentally partially matching different characters of a Chinese word is very low. This method takes advantage of the linguistic characteristics of Chinese compounds, which usually consist of two or three Chinese characters each representing the abbreviation of the original multi-character words. For example, the Chinese translations for the word *oral* and *exam* are 口的 *ko de* and 考試 *kaoshi*, respectively. Given that *oral* and *exam* are adjacent and that their Chinese translations match the first and second character of the word 口試 *koshi* in the Chinese text, they are most probably translations of 口試 if the positional difference between *oral exam* and 口試 happens to be small. Let us illustrate this method with concrete examples. Table 2 shows the information required to extract translations of Chinese compounds using this method.

Table 2. Extracted Translations of Chinese Compounds Based on Proximity

WI	English	WI	Chinese	MC
49	deep	26	根深蒂固	深
50	roots	26	根深蒂固	根
157	academic	112	學術界	學術
158	world	112	學術界	界
310	sexual	223	性騷擾	性
311	harassment	223	性騷擾	騷擾
501	eldest	387	長子	長
502	son	387	長子	子
1353	teaching	987	教材	教
1354	materials	987	教材	材
1886	oral	1412	口試	口
1887	exam	1412	口試	試
2443	research	1889	研究室	研究
2444	room	1889	研究室	室
3789	accept	3031	受辱	受
3790	insult	3031	受辱	辱

Based on Table 2, the equivalence between the following pairs can be identified and extracted *deep roots* ⇔ 根深蒂固 *genshengdigu*, *academic world* ⇔ 學術界 *xuesujie*, *eldest son* ⇔ 長子 *zhangzi*, *teaching materials* ⇔ 教材 *jiaocai*, *oral exam* ⇔ 口試 *koshi*, *research room* ⇔ 研究室 *yanjiushi*, *accept insult* ⇔ 受辱 *shouru*. This simple method can achieve more than 96% precision in extracting translations of Chinese compounds.

6. Using Co-occurrences of Word Pairs in Individual Documents to Filter Out Spurious Word Correspondences

Given that the precision of statistics-based and dictionary-based approaches is not satisfactory, is there any method of improving it? Melamed (1995) introduces a novel method to clean incorrect word correspondences induced by statistical word alignment algorithms that assign symmetrical association scores such as likelihood ratios. His approach is based on a sophisticated statistical model.

Instead of adopting document-internal distributional properties and the sophisticated statistical model proposed by Melamed, we tried to use some simple methods to improve the precision of translation lexicon derived from K-vec and dictionary lookup. We used the distributional information of the words in the individual documents that made up the corpus as a criterion of deciding which word pair was more likely to be correct. For instance, if *a* and *b* are an English-Chinese word pair, then there should be a significant correlation between the documents in which they appear.

To calculate the co-occurrences of a word pair in the whole corpus, each Chinese text was assigned the same index as its English translation. Two indexes, Chinese Word-Document index and English Word-Document index, were then constructed which recorded the document indexes of all the Chinese and English words occurring in the 58 Chinese and English documents in our corpus, respectively. The Jaccard Coefficient in (4) can be used as a measure to calculate the similarity of the distributions of a proposed English-Chinese word pair,

$$(4).\quad Jaccard(x,y) = \frac{c}{n_x + n_y - c}$$

where n_x and n_y are the number of documents in which x and y are found, and c is the number of documents they have in common.

The Jaccard Coefficient, however, is unreliable for high frequency words which appear in many documents.

In information retrieval, it has been established that the importance of a term is inversely proportional to the total number of documents in which it occurs (cf. Sparck Jones (1972)) In other words, the more documents a word occurs in, the less likely it is to be a keyword of the document. This measure, known as Inverse Document Frequency (IDF), is as follows.

$$(5)\quad \log_2 \frac{n}{DOCFREQ_k} + 1$$

where n is the total number of documents and $\mathrm{DOCFREQ}_k$ is the number of documents in which word *k* occurs.

We first used the output of dictionary lookup using partial matching and set the threshold of the Jaccard Coefficient to 0.5 and IDF to 2. After the filtering, we obtained a bilingual translation lexicon of 94% precision. Table 3 shows some of the results. The value of 0.5 of the Jaccard Coefficient was chosen in view of the fact that there are many morphological alternations of a given English word. The recall was about 43% for those word pairs that could be found by partial matching and 10% for the translations of all the English words in the text.

Table 3. Different Measures to Estimate the Co-occurrences of Word Pairs in Individual Documents Based on the Results of Dictionary Lookup

Chinese	English	Jaccard	C/N	E/N	IDF_C	IDF_E
傳統	traditional	0.71	0.81	0.85	2.46	2.54
中國	Chinese	0.67	0.97	0.69	1.86	1.37
根深蒂固	roots	0.153	1.00	0.15	5.86	3.16

In addition to the Jaccard Coefficient and IDF for Chinese and English words, we used two other ratios *C/N* and *E/N* to measure the distributional information, in which *N* is the number of common documents in which the proposed English and Chinese words in question occur and *c* and *e* are the number of documents in which the Chinese and English words occur. The information in the C/N and E/N columns provides further clues about the co-occurrence of a proposed word correspondence. It can be used to make inferences about cases where a word in the source text corresponds to more than one word in the target text. For instance, in Table 3, it shows that the word *roots* occurs in all the English corresponding documents where the word 根深蒂固 *genshengdigu* occurs. But of all the documents in which the word *root* occurs only 15.3% of the Chinese corresponding documents can find the occurrences of 根深蒂固. This suggests that the English word *root* might be part of the translation of the Chinese word 根深蒂固.

Although the method introduced above produced a high-precision lexicon, the recall was still low. This was partly because it could not adequately handle cases where a word in the source text corresponded to more than one word in the target text. This included cases where a Chinese word corresponded to an English lemma with several word forms (i.e. inflections) or cases where a Chinese compound or idiom corresponded to more than two English words. The first situation could be improved by lemmatisation, while the second situation could be improved by using proximity as discussed in the preceding section.

An alternative method to calculate the co-occurrences of word pairs in individual documents is to merge all the bilingual texts into one 'supper' bilingual text and then use K-vec again to estimate the co-occurrence of a word pair by regarding each document as a segment based on the output of the original K-vec or dictionary lookup. Table 4 shows the result of this method based on the word pairs extracted via the original K-vec.

Table 4. Using Co-occurrences of Word Pairs in the Whole Corpus to Clean the Output of the Original K-vec Algorithm

	Chinese	English	MI	t-score	Jaccard	Ratio_C	Ratio_E	IDF_C	IDF_E
+	媒體	media	1.60	2.68	0.84	0.88	0.94	2.68	2.77
+	教育	education	1.34	2.35	0.68	0.78	0.83	2.61	2.68
+	呂	Lu	2.20	2.07	0.58	0.87	0.63	3.85	3.39
%	不再	longer	1.14	1.97	0.54	0.76	0.65	2.77	2.53
+	權威	authority	2.31	1.95	0.54	0.85	0.60	4.05	3.53
+	劉	Liu	1.57	1.87	0.47	0.61	0.66	3.15	3.27
+	社會	society	0.62	1.85	0.75	0.87	0.84	1.85	1.81
+	張	Chang	0.54	1.43	0.56	0.67	0.77	1.90	2.10
%	性騷擾	harassment	4.85	1.36	1.00	1.00	1.00	5.85	5.27
%	性騷擾	sexual	4.27	1.34	0.66	1.00	0.66	5.85	5.27
+	導師	advisor	4.27	1.34	0.66	0.66	1.00	5.27	5.85

+	言論	speech	2.12	1.33	0.30	0.60	0.37	4.53	3.85
	亦	friend	1.05	1.26	0.30	0.50	0.42	3.27	3.05
	權	equal	2.53	1.17	0.28	0.50	0.40	4.85	4.53
	華	Kao	1.54	1.13	0.21	0.25	0.60	3.27	4.53
+	中國	Chinese	0.31	1.10	0.67	0.96	0.68	1.85	1.36
	便	increasingly	0.68	1.00	0.23	0.25	0.77	2.05	3.68
#	虹	Hsiaohung	5.85	0.98	1.00	1.00	1.00	6.85	6.85
%	校務	affairs	1.68	0.97	0.11	1.00	0.11	5.85	2.68
%	口試	oral	4.85	0.96	0.50	1.00	0.50	6.85	5.85
#	芬	Tehfen	4.85	0.96	0.50	0.50	1.00	5.85	6.85
+	友	friend	1.46	0.90	0.13	0.66	0.14	5.27	3.05
%	口試	exam	3.27	0.89	0.16	1.00	0.16	6.85	4.27
	使用	conference	1.05	0.89	0.17	0.21	0.50	3.05	4.27
	女性	male	1.27	0.82	0.16	1.00	0.16	6.85	4.27
	叫	Hsiaohung	2.39	0.81	0.09	0.09	1.00	3.39	6.85
	芬	Ho	2.27	0.79	0.14	0.50	0.16	5.85	4.27
	叫	name	0.45	0.71	0.22	0.63	0.25	3.39	2.10
#	德	Tehfen	1.68	0.68	0.05	0.05	1.00	2.68	6.85
	辯論	private	1.53	0.65	0.09	0.50	0.10	5.85	3.53
	民主	ethical	1.46	0.63	0.11	0.14	0.33	4.05	5.27
	權益	conference	1.27	0.58	0.11	0.25	0.16	4.85	4.27
	賀	each	0.49	0.58	0.09	1.00	0.09	4.85	1.50
+	事實	fact	0.18	0.49	0.37	0.72	0.43	2.39	1.64
+	平	equal	0.95	0.48	0.10	0.16	0.20	4.27	4.53
	性騷擾	another	0.53	0.43	0.05	1.00	0.05	5.85	1.53
	芬	each	0.49	0.41	0.04	1.00	0.04	5.85	1.50
	解嚴	recent	0.68	0.37	0.05	0.50	0.05	5.85	2.68
+	小	Hsiaohung	0.53	0.31	0.02	0.02	1.00	1.53	6.85
	根本	may	0.08	0.23	0.35	0.69	0.42	2.33	1.61
	德	Ho	0.10	0.09	0.09	0.11	0.33	2.68	4.27
	不再	authority	0.03	0.03	0.12	0.17	0.30	2.77	3.53
	也	with	0.00	0.00	0.94	1.00	0.94	1.07	1.00
	你	attack	-0.04	-0.06	0.11	0.13	0.50	1.95	3.85
	威權	generation	-0.12	-0.08	0.04	0.33	0.04	5.27	2.46
	德	each	-0.21	-0.51	0.22	0.61	0.26	2.68	1.50

+ Fully correct Pair   % The English gloss is part of the meaning of the Chinese gloss.
# The Chinese gloss is part of the meaning of the English gloss.

From Table 4, we can see that MI and t-score are more convenient than the Jaccard Coefficient, as the latter must be used in conjunction with IDF to filter out frequently occurring words. By setting the threshold of MI and t-score to 0 and 1.6, we extract 7 word pairs, all of which are correct, achieving 100% accuracy. However, the method leaves out 16 correct word pairs, obtaining a low recall of 0.3. Apparently, high precision is acquired at the cost of low recall. It is therefore more sensible to use this method to find anchor points. An interesting observation is that of all the 7 correct word pairs it identifies, only one of them involves collocations or compounds. The pair (不再 *buzai* ⇔ *longer*) is actually part of the translation (不再 *buzai* ⇔ *no longer*).

In the previous section, it has been shown that translations of Chinese compounds can be easily and accurately extracted using proximity condition. The proximity condition and the co-occurrences of word pairs in each document in the whole corpus therefore complement each other in extracting more anchor points.

7. Conclusion and Future Research

In this paper, we have shown the limitations of both statistics-based and dictionary-based approaches to deriving a translation lexicon. We have demonstrated the necessity of combining statistical and dictionary information and the usefulness of proximity and co-occurrence information of word pairs in individual documents in deriving a high-precision translation lexicon. Although the recall of our method is still low, its high precision can be used to find anchor points for deriving more word correspondences. We are currently implementing an iterative algorithm in the spirit of Kay and Röscheisen (1993) to improve the recall of the translation lexicon.. In the first iteration, the algorithm treats each individual document as a segment to calculate the co-occurrences of word pairs in the whole corpus. Word pairs which meet the proximity condition before the cleaning algorithm or above the threshold of co-occurrence ratio after the cleaning algorithm are identified as anchor points. These anchor points then become boundaries of segments for the next iteration. As the number of segments in each document increases, many correct word pairs incorrectly filtered out by the cleaning algorithm in the first few iterations can be expected to be recovered later.

Acknowledgement

The author wishes to express his gratitude to the *Sinorama* Magazine for permission to use the Sinorama Chinese-English parallel corpus and to Dr. Harold Somers, Dr. Tony McEnery, Mr. John McNaught, Prof. Chu.-Ren. Huang, Prof. Keh.-Jian Chen, Dr. Lee-Feng Chien and anonymous PACLIC reviewers for their helpful comments on an earlier draft of this paper.

References

Chen, K.-J. and Liu, S.-H. (1992) “Word Identification for Mandarin Chinese Sentences.” In Proceedings of the International Conference on Computational Linguistics, pp. 101-107.

Dagan, I., Church, W, and Gale, W. (1993) “Robust Bilingual Word Alignment for Machine Aided Translation. ”In Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, pp. 1-8, Ohio.

Dagan, I. (1996) “Bilingual Word Alignment and Lexicon Construction.” Tutorial paper given at the International Conference on Computational Linguistics, Copenhagen.

Fung, P. and Church, K. (1994) “K-vec: A New Approach for Aligning Parallel Texts.” Proceedings of the International Conference of Computational Linguistics, pp. 1096-1102, Kyoto.

Fung, P. and McKeown, K. (1994) “Aligning Noisy Parallel Corpora Across Language Groups: Word Pair Feature Matching by Dynamic Time Warping.” Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation, pp. 81-88.

Fung, P. and KcKeown, K. (1997) “A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora Across Language Groups.” Machine Translation, Vol. 12, Nos. 1-2., pp. 53-87.

Haruno, M. and Yamazaki, T. (1996) “High-Precision Bilingual Text Alignment Using Statistical and Dictionary Information.” Proceedings of Annual Conference of the Association for Computational Linguistics, pp. 131 – 138.

Jones, D. and Somers, H. (1995) “Bilingual Vocabulary Estimation from Noisy Parallel Corpora Using Variable Bag Estimation.” In JADT III GiornateInternazionali di Analsi Statistica dei Dati Testuali, pp. 255-262, Rome.

Kay, M. and Röscheisen, M. (1993) “Text-Translation Alignment.” Computational Linguistics, Vol. 19, No 1, pp 121-142.

Kumano, A. and Hirakawa, H. (1994) “Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistic Information.” in Proceedings of International Conference on Computational Linguistics, pp. 76-81, Kyoto.

Melamed, D. (1995) “Automatic Construction of Clean Broad-Coverage Translation Lexicons.” In Proceedings of 2nd Conference of the Association for Machine Translation in the Americas, Montreal.

Somers, H. and Ward, A. (1996) “Some More Experiments in Bilingual Text Alignments.” In Oflazer, K. and Somers, H. (eds.) Proceedings of the Second International Conference on New Methods in Language Methods in Language Processing, pp. 66-78, Ankara.

Sparck Jones, K. (1972) “A Statistical Interpretation of Term Specificity and Its Application in Retrieval. Journal of Documentation, Vol. 28, No. 1, pp. 11-21.

Utsuro, T. et al. (1994) “Bilingual Text Matching Using Bilingual Dictionary and Statistics.” in Proceedings of International Conference on Computational Linguistics, pp. 1076-1082, Kyoto.

Wu, D. and Xia, X. (1995) “Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon.” Machine Translation, Vol. 9, pp. 285-313.