TopAlign: Word Alignment for Bilingual Corpora based on Topical Clusters of Dictionary Entries and Translations

Mathis H. Chen, Jason S. Chang, Sue J. Ker and Jen-Nan Chen
Department of Computer Science
National Tsing Hua University
Hsinchu 30043, Taiwan, R.O.C.
mathis@nlplab.cs.nthu.edu.tw, jschang@cs.nthu.edu.tw,
ksj@volans.cis.scu.edu.tw, dr828310@nlplab.cs.nthu.edu.tw

Abstract. *Aligned parallel corpora have proved very useful in many natural language processing (NLP) tasks, including statistical machine translation and word sense disambiguation. In this paper, we address major issues relating to current research in word alignment: language-independence and broad lexical coverage. In addressing these issues, we will discuss the central problems of data sparseness and noise in knowledge acquisition and suggest an approach based on a bilingual machine readable dictionary (MRD). We will describe an MRD-based method called TopAlign for word alignment which relies on topical clustering of dictionary entries including headwords and translations. While not requiring a very large bilingual corpus, the language-independent approach underlaying TopAlign rivals corpus-based methods for coverage as well as precision.*

1. Introduction

Aligned corpora have proved very useful in many tasks, including statistical machine translation (SMT) [Brown et al., 1990; Wu & Ng, 1995] and word sense disambiguation [Chang et al., 1996]. Several methods have recently been proposed for sentence alignment of the *Hansards*, an English-French corpus of Canadian parliamentary debates [Brown et al., 1991; Gale & Church, 1991a; Simard et al., 1992; Chen, 1993; Gale & Church, 1993]; and for other language pairs, including English-German, English-Chinese, and English-Japanese [Kay & Röscheisen, 1993; Church et al., 1993; Chang & Chen, 1997].

The SMT approach can be understood as a word-by-word model consisting of two sub-models: a *language model* for generating a source text segment S and a *translation model* for mapping S to its translation T. Brown et al. [1990] recommend using a bilingual corpus to train the parameters of *translation probability*, $\Pr(S|T)$ in the translation model. In the context of SMT, Brown, Della Pietra, Della Pietra, and Mercer [1993] present a series of five models of $\Pr(S|T)$ for word alignment. They propose using an adaptive *Expectation and Maximization* (EM) algorithm to estimate parameters for $\Pr(S|T)$ from a bilingual corpus. The EM algorithm iterates between two phases to estimate the two factors of $\Pr(S|T)$, namely *lexical translation probability* (LTP) and *distortion probability* (DP) until both functions converge. The SMT model is then tested for the task of machine translation. The model produces thirty-five acceptable English translations for seventy-three French sentences.

For efficient alignment of parallel text written in any language pairs, we believe that the following questions should be asked:

1. Is the method language-pair independent?
2. Does the method provide global coverage of bilingual lexical mappings?
3. Does the method provide precise bilingual lexical mapping?

Dagan, Church and Gale [1993] observe that reliably distinguishing sentence boundaries for noisy bilingual texts obtained from OCR (optical character recognition) devices is quite difficult. The

authors recommend aligning words directly without the preprocessing phase of sentence alignment. They observe that there are many instances of cognates among the languages in the Indo-European family. Based on the observation, a rough char-based alignment is performed first, which provides a base for estimating the translation probability based on position, as well as limits the range of alignment target. However, Fung and Church [1994] point out that such a cognate-based constraint does not exist in pairs across language groups such as Chinese and English. The authors propose a *K-vec* approach which is based on a *k*-way partitioning of the bilingual corpus. Fung and McKeown [1994] propose using a similar measure based on *Dynamic Time Warping* (DTW) between occurrence recency sequences to improve on the *K-vec* method.

K-vec, *DK-vec*, *DTW*, and many other proposed methods for word alignment and bilingual lexicon construction, including ϕ^2 [Gale & Church, 1991b], cognates [Brown et al., 1991; Simard et al., 1992], are based primarily on co-occurrences of words and translations. Gale and Church [1991b] and Macklovitch and Hannan [1996] point out that co-occurrence-based statistics are very unreliable for situation where the data samples are sparse. Although Brown et al.'s Model 1 converges in the course of iterative refinement, the other four models may not bear the same property. This is aggravated by the fact that word-based co-occurrence statistics estimated from a bilingual corpus are found to be seriously faulted in terms of precision. Macklovitch and Hannan [1996] suggest that richer, more abstract representations are required in order to provide broad and precise methods for the ultimate goal of translation analysis. Knowledge acquired at a more abstract level such as genus terms in MRD and synonyms in a thesaurus category are beginning to be exploited to cope with the robustness and data sparseness issues in problems ranging from noun sequence interpretation [Vanderwende, 1994] to word sense disambiguation [Yarowsky 1992], and to word alignment [Ker & Chang, 1997].

This paper describes an MRD-based method called *TopAlign* for word alignment which relies on topical clustering of dictionary entries including headwords and translations to provide estimates of LTP. While not requiring a very large bilingual corpus, *TopAlign* rivals corpus-based methods for coverage as well as precision. Furthermore, the approach only requires widely available resources such as a bilingual dictionary and a source-language thesaurus. Therefore, *TopAlign* does not rely on language specific properties, therefore is, to some extend, language independent.

2. Diverse In-Context Translation and Robust Estimation of LTP

A wide variety of ways of estimating the lexical translation probability (LTP) have been proposed in the literature of computational linguistics. However, the experimental results indicate that we are still left without a simple, straightforward method that can cope with diverse translations.

Given that dictionary translations (DT) for headwords can be extracted from a bilingual *machine readable dictionary* (MRD) such as the Longman English-Chinese Dictionary of Contemporary English (LecDOCE), words can be easily aligned with their translations based on DTs. Headword-and-translation pairs are a reliable knowledge source for word alignment, resulting in highly precise connections (over 95%) for bilingual LecDOCE examples and texts of a computer manual. Ker and Chang [1997] report that the translations of a word in context (*In-Context Translation*, ICT for short) are frequently more diversified than the offering in an everyday bilingual dictionary. More specifically, less than 30% of the English words in the context of an LecDOCE example translate into one of the relevant DTs in the same dictionary. A probabilistic lexicon derived from a very large bilingual corpus fares much better but still covers just over 60% of the lexical mappings required for complete alignment. The low coverage should not come as a surprise, as we will show below that ICTs are very diverse thus lack distributional regularity.

Now, let us look at the diversity of ICTs more closely. A translation in an everyday dictionary is meant to provide the reader with an idea of what is implied by the headword out of context. DTs are frequently more of an explanation rather than translation. Aside from this fundamental difference, the

disparity between a DT and an ICT may arise due to many linguistic phenomena.

Transformation of Part-of-Speech. A source word of a certain syntactic category might translate to a target word of a different category, leading to diverse translations not listed in a bilingual MRD. The translations "下落" of "*happen*" in Example 1 is not among "發生," "碰巧," and "偶然發生," the *happen*-DTs listed in the LecDOCE. This is to be expected since the part-of-speech of "*happen*" changes from verb to noun in the course of translation to suit the particular context of Example 1.

Example 1 There's something funny about this affair; no one seems to know what's happened to all the money. 這件事有點可疑，似乎誰都不知道這些錢的下落。

Sense Gaps or Sense Shifts. Novel sense shifts are sometimes too dynamic and numerous for a dictionary to cover exhaustively. The translation of "*click*" in Example 2 seems to indicates that the sense of making a clicking sound shifts toward the action "按" ("*press*") which cause the sound. Such a shifted sense is absent from the LecDOCE.

Example 2 Click the mouse button. 按滑鼠鍵。

Collocations. A collocation is an arbitrary and recurrent word combination [Smadja, 1993]. In most cases, collocations do not translate in a word by word fashion. For instance, the word "*lose*" in either phrase "*lose one's way*," or "*lose one's car*" bears the same meaning "*fail to find.*" However, the translations are "迷" and "丟", respectively, which are not interchangeable; one never says "丟了路" or "迷了車" in Mandarin. Collocations lead to very diverse ICTs, especially for light verbs such as "*lose*," "*put*," "*get*," etc.

Example 3 He lost his way in the mist. 他在霧中迷了路。 (*他在霧中丟了路。)

Interchangeable Synonymous Translations. Finally, the diversity of translation can be attributed also to interchangeable synonymous translations. For instance in Example 4, "遇見" the translation of "*meet*" is synonymous to the relevant DT "遇到."

Example 4 I have never met so nice a girl. 我從未遇見過這麼好的女孩子。

3. A Lexical Translational Representation with Broad Coverage

The way to cope with a plethora of diverse ICTs must obviously come from a richer and more abstract representation, which provides classification of words or word senses to bound ICTs. What is needed is a classification that is independent of language, part-of-speech, argument structure, etc. To derive an ideal classification immediately faces the problem of knowledge acquisition bottleneck. It is, therefore, tempting to exploit abstract representations readily available in lexicographic resources such as the Roget's thesaurus, the WordNet, and the Longman Dictionary of Contemporary English (LLOCE) [McArthur, 1992]. Close examination seems to indicate that the topical classification of word senses is well suited for this particular NLP task of word alignment. The diverse ICTs seem to be constrained, by and large, within the translations of some topical cluster of word senses. For instance, the LecDOCE examples reproduced as Example 5 and 6 indicate that the ICTs, "遇見" and "遇到" of words "meet" and "encounter" list under Mc072 (Meeting people and things) in the LLOCE, are also the DTs of other Mc072-words.

Example 5 I have never met so nice a girl. 我從未遇見這麼好的女孩子.

Example 6 He encountered many difficulties. 他遇到很多困難.

Furthermore, such ICTs and DTs are often synonymous compounds that share a common morpheme; in this case, the morpheme "遇." Example 7 indicates such morpheme-sharing synonymous compounds exist among many (ICT, DT) pair of a source word. For instance, the pair of translations (女士, 女性) share a common morpheme "女" (woman). Fujii and Croft [1993] point out

a similar thesaurus effect of Chinese morpheme in the context of Japanese information retrieval.

Example 7 She's a very wealthy woman, and moves in the highest circles of society.

她是位很有錢的女士，活躍於高級社交圈. （女性∈ DT_{woman}）

These observations suggest that LTP can be estimated more robustly via cluster-to-morpheme mapping. We have adopted the *TopSense* method proposed by Chen and Chang [1997] to cluster each LecDOCE entries to one of the topical sense clusters (TSC) and topical translation clusters (TTC) in the LLOCE. Based on a 12-word evaluation, *TopSense* clusters 93% of dictionary senses with a precision rate of 92%.

Estimation of Topic-based Word-to-Morpheme LTP

Armed with TSC and TTC, we define the relevance of a morpheme m to a topic t in terms of a weight $w(t, m)$. Such weights can be obtained in a manner similar to what is done in IR when assigning weights to index terms. The relevancy of a source word s to a translation morpheme m, $R(s, m)$, is given by the following equations:

$$R_{TTC}(s, m) = \max_{t \in TOP_s} w(t, m) \quad \text{(Eq. 1)}$$

$$w(t, m) = tf_{t,m} \times idf_m \quad \text{(Eq. 2)}$$

$$idf_m = \log \frac{T}{df_m} \quad \text{(Eq. 3)}$$

where $tf_{t,m}$ = the frequency of the morpheme m appearing in the topical translation cluster t,
T = the total number of TTCs,
df_m = the number of TTCs that contain the morpheme m, and
TOP_s= the set of TTCs to which a word sense of s belongs.

This relevancy score is intended to compensate what is lacking in the offering in a bilingual dictionary to arrive at a broad-coverage and precise estimate of the LTP.

Estimation of DT-based Word-to-morpheme LTP

We use the word to morpheme LTP, $t(s, m)$ to denote how likely an English word s translates to words containing the morpheme m. We have adopted a statistical estimator based on likelihood ratio to estimate the LTP. The estimator, $R_{DT}(s, m)$, is given by the following equations:

$$R_{DT}(s, m) = \frac{Pr(m|s)/Pr(\neg m|s)}{Pr(m|\neg s)/Pr(\neg m|\neg s)}, \quad \text{(Eq. 4)}$$

$$Pr(m|s) \cong \frac{|FROM_s \cap TO_m|}{|FROM_s|}, \quad \text{(Eq. 5)}$$

$$Pr(m|\neg s) \cong \frac{|(TRANS - FROM_s) \cap TO_m|}{|TRANS - FROM_s|}, \quad \text{(Eq. 6)}$$

Where $Pr(\neg m|s) = 1 - Pr(m|s)$
$Pr(\neg m|\neg s) = 1 - Pr(m|\neg s)$
TRANS = the set of all dictionary translations,
$FROM_s$ = the set of dictionary translations for a given source word s,
To_m = the set of dictionary translations containing the morpheme m.

$Pr(m|s)$ is the probability of translation of s contains the morpheme m. However, MLE estimation of $Pr(m|s)$ would assign zero probability to all unseen data. Furthermore, it fails to provide a reliable estimate for cases where the data is sparse. To resolve this problem, we smooth zero frequencies by

assigning a small probability value empirically determined.

Estimation of LTP based on R_{DT} and R_{TTC}

In the following, we describe a word to morpheme LTP that combines dictionary-based estimates with topic-based estimates. The LTP t(s, m) is defined by the following cases:

Case 1. $R_{DT}(s, m) \geq h_1$,
Case 2. $h_1 > R_{DT}(s, m) \geq h_2$,
Case 3. $h_2 > R_{DT}(s, m) \geq h_3$ and $R_{TTC}(s, m) \geq h_4$
Case 4. $h_2 > R_{DT}(s, m) \geq h_3$ and $R_{TTC}(s, m) < h_4$
Case 5. $R_{DT}(s, m) < h_3$ and $R_{TTC}(s, m) \geq h_4$
Case 6. $R_{DT}(s, m) < h_3$ and $R_{TTC}(s, m) < h_4$.

The connections satisfying each condition are given the same probability value determined by maximal likelihood estimation (MLE). For instance, if there are k connections in a sample of n candidates (s, m) such that t(s, m) $\geq h_1$ then all these candidates are given the same MLE value for LTP, i.e. t(s, m) = t_1 = k/n. Equation (Eq. 7) sums up the above discussion:

$$t(s,m) = \begin{cases} t_1 \text{ if } R_{DT}(s,m) \geq h_1, \\ t_2 \text{ if } h_1 > R_{DT}(s,m) \geq h_2, \\ t_3 \text{ if } h_2 > R_{DT}(s,m) \geq h_3 \text{ and } R_{TTC}(s,m) \geq h_4, \\ t_4 \text{ if } h_2 > R_{DT}(s,m) \geq h_3 \text{ and } R_{TTC}(s,m) < h_4, \\ t_5 \text{ if } R_{DT}(s,m) < h_3 \text{ and } R_{TTC}(s,m) \geq h_4, \\ \tau_6 \text{ if } R_{DT}(s,m) < h_3 \text{ and } R_{TTC}(s,m) < h_4. \end{cases} \quad \text{(Eq. 7)}$$

By using a small sample of a few hundred sentences, the LTP values t_i for $1 \leq i \leq 6$ can be estimated as described. Table 1 summarizes the MLE probabilistic values associated with lexical and conceptual factors estimated using 200 sentences from the LecDOCE.

Conceptual and Lexical Conditions	# Candidates	# Connections	MLE of t(s, t)	
$R_{DT}(s, m) \geq 100$	1158	966	t_1	0.834
$100 > R_{DT}(s, m) \geq 10$	1318	659	t_2	0.500
$10 > R_{DT}(s, m) \geq 5$ and $R_{TTC}(s, m) \geq 10$	429	134	t_3	0.312
$10 > R_{DT}(s, m) \geq 5$ and $R_{TTC}(s, m) < 10$	502	77	t_4	0.153
$R_{DT}(s, m) < 5$ and $R_{TTC}(s, m) \geq 10$	8977	853	t_5	0.095
$R_{DT}(s, m) < 5$ and $R_{TTC}(s, m) < 10$	13809	428	t_6	0.031

Table 1. Maximum likelihood estimation (MLE) of LTP for word-to-morpheme LTP.

4. The *TopAlign* Algorithm

As mentioned earlier, Brown Mode 2 stipulates that a connection be given a probability value Pr(s, t) as the product of LTP, t(s | t) and DP, d(i | j, l, m). Corresponding to the model, we also give each connection candidate a probabilistic value according to lexical and position considerations. For the estimation of DP, we adopt the convolution-based method proposed by Chang and Chen (1997). Under the method, the distortion probability becomes a (2w+1) by (2w+1) array, mask(i, j) used in 2-dimensional discrete convolution operation, a popular techniques in IP. We give each connection candidate a probabilistic value according to (Eq. 8):

$$\Pr(s_x, t_y) = \sum_{j=-w}^{w} \sum_{i=-w}^{w} t(s_{x+i}, t_{y+j}) \times \text{mask}(i, j) \quad \text{(Eq. 8)}$$

where w is a pre-determined parameter specifying the size of the convolution filter. Connections that fall outside this window are assumed to have no affect on $\Pr(s_x, t_y)$. The above description of word alignment is summarized as the *TopAlign* algorithm.

Algorithm (*TopAlign*) TTC-based word alignment for a pair of sentences (S, T)

Step 1: Initialize the result *ANS* to an empty list.

Step 2: Tag each word in S with POS information and convert each word to the root form to obtain the sequence W_S of words in S.

Step 3: Calculate the relevancy of a source word s to a translation morpheme m, $t(s, m)$, according to (Eq. 1) through (Eq. 7) for each s in W_S and m in T.

Step 4: For each connection candidate (s_x, t_y), $s_x \in S$, $t_y \in T$, compute $\Pr(s_x, t_y)$ according to Equations (Eq. 8).

Step 5: Add to *ANS* the connection (s_x^*, t_y^*) that maximizes $\Pr(s_x, t_y)$ over all $s_x \in S$, $t_y \in T$ with a value greater than h. This step repeats itself until candidates run out or all candidate (s_x, t_y) is associated with a $\Pr(s_x, t_y)$ value lower than h.

Step 6: Output *ANS* as the final result of word alignment.

An Illustrative Example. In the following, we demonstrate how to estimate the class-based word-to-morpheme LTP using Example 8.

Example 8 The$_1$ old$_2$ lady$_3$ was$_4$ clad$_5$ in$_6$ a$_7$ fur$_8$ coat$_9$.$_{10}$

這$_1$位$_2$老$_3$婦人$_4$穿著$_5$皮$_6$裘$_7$。$_8$

	這	位	老	婦	人	穿	著	皮	裘
the	0.031	0.031	0.031	0.031	0.031	0.031	0.031	0.031	0.031
old	0.031	0.031	0.500	0.031	0.031	0.031	0.031	0.031	0.031
lady	0.031	0.500	0.031	0.834	0.500	0.031	0.031	0.031	0.031
be	0.031	0.031	0.031	0.031	0.031	0.031	0.031	0.031	0.031
clad	0.031	0.031	0.031	0.095	0.031	0.834	0.834	0.031	0.031
in	0.031	0.031	0.031	0.031	0.031	0.500	0.153	0.031	0.031
a	0.031	0.031	0.031	0.031	0.031	0.031	0.031	0.031	0.031
fur	0.031	0.031	0.031	0.031	0.031	0.031	0.031	0.834	0.031
coat	0.031	0.031	0.031	0.095	0.031	0.095	0.031	0.500	0.031

Table 2. The connection candidates (s, m) in Example 9 with LTP values $t(s, m)$.

	這	位	老	婦	人	穿	著	皮	裘
the	0.082	-0.071	-0.287	-0.320	-0.262	-0.143	-0.065	0.026	0.060
old	-0.037	0.112	0.582	-0.052	-0.321	-0.186	-0.106	0.003	0.037
lady	0.077	0.407	0.048	0.687	0.760	-0.112	-0.320	-0.156	-0.053
be	-0.168	-0.340	-0.375	-0.267	-0.141	-0.131	-0.267	-0.226	-0.083
clad	-0.080	-0.213	-0.267	-0.239	-0.139	0.498	0.661	0.020	-0.099
in	0.008	-0.028	-0.059	-0.152	-0.142	0.073	-0.072	-0.097	-0.022
a	0.008	-0.026	-0.063	-0.143	-0.242	-0.249	-0.270	-0.247	-0.116
fur	0.031	0.011	-0.023	-0.046	-0.061	-0.130	0.051	0.432	0.119
coat	0.048	0.020	0.014	0.054	0.036	-0.083	-0.032	0.190	0.137

Table 3. Pr (s, t) values for Example 9 after applying the distortion filter to the LTP values. Notice that the filter does successful remove the noise of (in, 穿), (in, 著), and (coat, 皮), and enhance the missing signal for the connecting (coat, 裘).

Table 2 lists all connection candidates along with their LTP $t(s, m)$ values. After applying the convolution filter to the LTP values, the Pr (s, t) values for Example 9 are as shown in Table 3. With all these calculations done after executing Step 4, *TopAlign* selects the highest pr(s, t) candidate to put

into *ANS*. *TopAlign* stops after running out of connections with a probabilistic value greater than *h*, 0.05. The success rate is evaluated according to how many English words are correctly aligned. Evaluation is based on 100% coverage, i.e. each word in the English sentence is checked for correct alignment. A word not given a connection is considered a failure if it should be connected to some Mandarin morpheme *s*; otherwise it is considered a success. For this example, 9 of 9 are aligned correctly. Therefore, the success rate is 9/9 = 100%.

5. Experimental Results

The proposed method's effectiveness, we have implemented the algorithms described in Section 3 and conducted a series of experiments. Tests are performed on the sentences found in the LecDOCE and bilingual computer manuals to assess the method's robustness and generality. One of the test sets consists of 200 examples and their Mandarin translations randomly extracted from the LecDOCE. The English example sentences range from 8 to 23 words long. Average sentence length is 11.5 words. The evaluation shows that over 80% of the English words are aligned at a high precision rate of around 90%. The effectiveness of *TopAlign* is due mostly to robust estimation of LTP based on topical clusters of word senses and translation morphemes.

6. Discussion and Concluding Remarks

This paper has presented a simple and fast method capable of identifying words and their in-context translation in a bilingual corpus. The proposed algorithm produces broad and precise connections for specific linguistic reasons. A typical source word tends to have diversified translations infrequently found in a bilingual MRD or corpus-derived probabilistic lexicon. However, we observe that these diverse translations are, by and large, bounded within the topical translation cluster for the relevant word sense. For mono-syllabic languages such as Mandarin which are often rich in compounding, a further constraint related to synonymous compounding morpheme can be exploited to further lower the complexity of the search for the ICTs. In view of these, we contend that an approach based on topic clusters of dictionary senses and translation morpheme can address the issues raised in the Introduction. Our assumption seems to hold out since the experiments in this study demonstrate that the method provides broad-coverage as well as high-precision alignment.

References

[Brown et al., 1991] Brown, P. F., J. C. Lai and R. L. Mercer, 1991. Aligning Sentences in Parallel Corpora, In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 169-176, Berkeley, CA, USA.

[Brown et al., 1990] Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roosin, 1990. A Statistical Approach to Machine Translation, *Computational Linguistics*, 16:2, 79-85.

[Chang et al., 1996] Chang, J. S., J. N. Chen, H. H. Sheng and S. J. Ker, 1996. Combining Machine Readable Lexical Resources and Bilingual Corpora for Broad Word Sense Disambiguation, In *Proceedings of the second Conference of the Association for Machine Translation in the Americas*, 115-124, Montreal, Canada.

[Chang & Chen, 1997] Chang, J. S. and M. H. Chen, 1997. An Alignment Method for Noisy Parallel Corpora based on Image Processing Techniques, To appear in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain.

[Chen & Chang, 1997] Chen, J. N. and J. S. Chang, TopSense: A Topical Sense Clustering Method Based on Information Retrieval Techniques on Machine Readable Resources, To appear In *Special Issue on Word Sense Disambiguation, Computational Linguistics*.

[Chen, 1993] Chen, S. F., 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information, In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 9-16, Columbus, Ohio, USA.

[Church et al., 1993] Church, K. W., I. Dagan, W. A. Gale, P. Fung, J. Helfman, and B. Satish, 1993. Aligning Parallel Texts: Do Methods Developed for English-French Generalize to Asian Languages? In *Proceedings of the First Pacific Asia Conference on Formal and Computational Linguistics*, 1-12, Taipei, Taiwan.

[Fujii & Croft, 1993] Fujii, H. and W. B. Croft, 1993. A Comparison of Indexing Techniques for Japanese Text Retrieval, In *Proceedings of the 6th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 237-247, Pittsburgh, PA, USA.

[Fung & Church, 1994] Fung P. and Church K. W. , 1994. K-vec: A New Approach for Aligning Parallel texts. In *Proceedings of the 15th International Conference on Computational Linguistics*, 1096-1102, Kyoto, Japan.

[Fung & McKeown, 1994] Fung, P. and K. McKeown, 1994. Aligning Noisy Parallel Corpora across Language Groups: Word Pair Feature Matching by Dynamic Time Warping, In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 81-88, Columbia, Maryland, USA.

[Gale & Church, 1991a] Gale, W. A. and K. W. Church, 1991a. A Program for Aligning Sentences in Bilingual Corpora, In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 177-184, Berkeley, CA, USA.

[Gale & Church, 1991b] Gale, W. A. and K. W. Church, 1991b. Identifying Word Correspondences in Parallel Texts, In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, 152-157, Pacific Grove, CA, USA.

[Gale & Church, 1993] Gale, W. A. and K. W. Church, 1993. A Program for Aligning Sentences in Bilingual Corpora, *Computational Linguistics*, 19:1, 75-102.

[Kay & Röcheisen, 1993] Kay, M. and M. Röcheisen, 1993. Text-Translation Alignment, *Computational Linguistics*, 19:1, 121-142.

[Ker & Chang, 1997] Ker, S. J, and J. S. Chang, 1997. A Class-base Approach to Word Alignment, *Computational Linguistics*, Vol. 23, No. 2, pp. 313-343, June.

[Macklovitch & Hannan, 1996] Macklovitch, E. and Hannan M-L., 1996. Line 'Em Up: Advances In Alignment Technology and Their Impact on Translation Support Tools, In *Proceedings of the second Conference of the Association for Machine Translation in the Americas (AMTA)*, 145-156, Montreal, Canada.

[McArthur, 1992] McArthur, T., 1992. *Longman Lexicon of Contemporary English*, Published by Longman Group (Far East) Ltd., Hong Kong.

[Simard et al., 1992] Simard, M., G. F. Foster, and P. Isabelle, 1992. Using Cognates to Align Sentences in Bilingual Corpora, In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, 67-81, Montreal, Canada.

[Smadja, 1993] Smadja, F., 1993. Retrieval Collocations from Text: Xtract. *Computational Linguistics*, 19:1, 143-178.

[Vanderwende, 1994] Vanderwende, L., 1994. Algorithm to Automatic Interpretation of Noun Sequences, In *Proceedings of the 15th International Conference on Computational Linguistics*, 782-788, Kyoto, Japan.

[Wu & Ng, 1995] Wu, D. and C. Ng, 1995. Using Brackets to Improve Search for Statistical Machine Translation. In *Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation*, 195-204, Kowloon, Hong Kong.

[Yarowsky, 1992] Yarowsky, D., 1992. Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, In *Proceedings of the 14th International Conference on Computational Linguistics*, 454-460, Nantes, France.