Aligning More Words with High Precision for Small Bilingual Corpora

Sur-Jin Ker
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan, ROC 30043
jschang@cs.nthu.edu.tw

Jason J. S. Chang
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan, ROC 30043
jschang@cs.nthu.edu.tw

Abstract

In this paper, we propose an algorithm for aligning words with their translation in a bilingual corpus. Conventional algorithms are based on word-by-word models which require bilingual data with hundreds of thousand sentences for training. By using a word-based approach, less frequent words or words with diverse translations generally do not have statistically significant evidence for confident alignment. Consequently, incomplete or incorrect alignments occur. Our algorithm attempts to handle the problem using class-based rules which are automatic acquired from bilingual materials such as a bilingual corpus or machine readable dictionary. The procedures for acquiring these rules is also described. We found that the algorithm can align over 80% of word pairs while maintaining a comparably high precision rate, even when a small corpus was used in training. The algorithm also poses the advantage of producing a tagged corpus for word sense disambiguation.

1. Introduction

Brown et al. (1990) initiated much of the recent interest in bilingual corpora. They advocated applying a statistical approach to machine translation (SMT). The SMT approach can be understood as a word by word model consisting of two submodels: a language model for generating a source text segment ST and a translation model for translating ST to a target text segment TT. They recommended using an aligned bilingual corpus to estimate the parameters of translation probability, Pr(ST |TT) in the translation model. The resolution of alignment can vary from low to high: section, paragraph, sentence, phrase, and word (Gale and Church 1993; Matsumoto et al. 1993).

In addition to machine translation, many applications for aligned corpora have been proposed, including bilingual lexicography (Gale and Church 1991, Smadja 1992, Dallie, Gaussier and Lange 1994), and word-sense disambiguation (Gale, Church and Yarowsky 1992, Chen and Chang 1994).

In the context of statistical machine translation, Brown et al. (1993) presented a series of five models for Pr(ST |TT). The first two models have been used in research on word alignment. Model 1 assumes that Pr(ST|TT) depends only on lexical translation probability t(s | t), i.e., the probability of the i-th word in ST producing the j-th word t in TT as its translation. The pair of words (s, t) is called a connection. Model 2 enhances Model 1 by considering the dependence of Pr(ST |TT) on the distortion probability, d(i | j, l, m) where l and m are the numbers of words in ST and TT, respectively.

Using an EM algorithm for Model 2, Brown et al. (1990) reported the model produced seventeen acceptable translations for twenty-six testing sentences. However, the degree of success in word alignment was not reported.

Dagan, Church and Gale (1992) proposed directly aligning words without the preprocessing phase of sentence alignment. Under this proposal, a rough character-by-character alignment is first performed. From this rough character alignment, words are aligned using an EM algorithm for Model 2 in a fashion quite similar to the method presented by Brown. Instead of d(i | j, l, m), a smaller set of offset probabilities, o(i - i') were used where the i-th word of ST was connected to the j-th word of TT in the rough alignment. This algorithm was evaluated on a noisy English-French technical document. The authors claimed that 60.5% of 65,000 words in the document were correctly aligned. For 84% of the words, the offset from correct alignment was at most 3.

Motivated by the need to reduce on the memory requirement and to insure robustness in estimation of probability, Gale and Church (1991) proposed an alternative algorithm in which probabilities are not estimated and stored for all word pairs. Instead, only strongly associated word pairs are found and stored. This is achieved by applying ϕ^2 test, a χ^2-like statistic. The extracted word pairs are used to match words in ST and TT. The algorithm works from left to right in ST, using a dynamic programming procedure to maximize Pr(ST |TT). The probability t(s | t) is approximated as a function of fan-in, the number of matches (s', t) for all s' $\in$ ST, while distortion d(i | j, l, m) is approximated as a probability function, Pr(match|j'-j) of slope, j'-j, where (i', j') is the positions of the nearest connection to the left of s. The authors claim that when a relevant threshold is set, the algorithm can recommend connections for 61% for

the words in 800 sentence pairs. Approximately 95% of the suggested connections are correct.

In this paper, we propose a word-alignment algorithm based on classes derived from sense-related categories in existing thesauri. We refer to this algorithm as SenseAlign. The proposed algorithm relies on an automatic procedure to acquire class-based rules for alignment. It does not employ word-by-word translation probabilities; nor does it use a lengthy iterative EM algorithm for converging to such probabilities. Results obtained from the algorithms demonstrate that classification based on existing thesauri is very effective in broadening coverage while maintaining high precision. When trained with a corpus only one-tenth the size of the corpus used in Gale and Church (1991), the algorithm aligns over 80% of word pairs with comparable precision (93%). Besides, since the rules are based on sense distinction, word sense ambiguity can be resolved in favor of the corresponding senses of rules applied in the alignment process.

The rest of this paper is organized as follows. In the next section, we describe SenseAlign and discuss its main components. Examples of its output are provided in Section 3. All examples and their translations are taken from the Longman English-Chinese Dictionary of Contemporary English (Procter 1988, LecDOCE, henceforth). Section 4 summarizes the results of inside and outside tests. In Section 5, we compare SenseAlign to several other approaches that have been proposed in literature involving computational linguistics. Finally, Section 6 summarized the paper.

2. The Word Alignment Algorithm

2.1 Preliminary details. SenseAlign is a class-based word alignment system that utilizes both existing and acquired lexical knowledge. The system contains the following components and distinctive features.

A. **A greedy algorithm for aligning words.** The algorithm is a greedy decision procedure for selecting preferred connections. The evaluation is based on composite scores of various factors: applicability, specificity, fan-out, relative distortion probabilities, and evidence from bilingual dictionaries.

B. **Lexical preprocessing.** Morphological analysis, part-of-speech tagging, idioms identification are performed for the two languages involved. In addition, certain morpho-syntactic analyses are performed to handle structures that are specific only to one of the two languages involved. By doing so, the sentences are brought closer to each other in the number of words.

C. **Two thesauri for classifying words.** (McArthur 1992; Mei et al. 1993) Classification allows a word to align with a target word using the collective translation tendency of words in the same class. Class-base rules obviously have much less parameters, are easier to acquire and can be applied more broadly.

D. **Two different ways of learning class-based rules.** The class-based can be acquired either from bilingual materials such as example sentences and their translations or definition sentences for senses in a machine readable dictionary.

E. **Similarity between connection target and dictionary translations.** In 40% of the correct connections, the target of the connection and dictionary translation have at least one Chinese character in common. To exploit this thesaury[1] effect in translation, we include similarity between target and dictionary translation as one of the factors.

F. **Relative distortion.** Translation process tends to preserve contiguous syntactical structures. The target position in a connection high depends that of adjacent connections. Therefore, parameters in an model of distortion based on absolute position are highly redundant. Replacing probabilities of the form d(i|j, l, m) with relative distortion is a feasible alternative. By relative distortion, rd for the connection (s,t), we mean (j-j')-(i-i') where i'th word, s' in the same syntactical structure of s, is connected to the j'th word, t' in TT.

2.2. Acquisition of alignment rules. Class-based alignment rules can be acquired from a bilingual corpus. Table 1 presents the ten rules with the highest applicability acquired from the example sentences and their translations in LecDOCE. Alternatively, we can acquire rules from the bilingual definition text for senses in a bilingual dictionary. The definition sentence are disambiguated using a sense division based on thesauri for the two language involved. Each sense is assigned codes from the two thesauri according to its definition in both languages. See Table 2 for examples of sense definition and acquired rules.

2.3 Evaluation of connection candidates. Connection candidates can be evaluated using various factors of confidence. The probabilities of having a correct connection as functions of these factors are estimated empirically to reflect their relative contribution to the total confidence of a connection

[1] From one aspect, those words sharing common characters can be considered as synonyms that would appear in a thesaurus. Fujii and Croft (1993) pointed out that this thesaury effect of Kanji in Japanese helps broaden the query favorably for character-based information retrieval of Japanese documents.

candidate. Table 3 lists the empirical probabilities of various factors.

2.4. Alignment algorithm. Our algorithm for word alignment is a decision procedure for selecting the preferred connection from a list of candidates. The initial list of selected connection contains two dummy connections. This establishes the initial anchor points for calculating relative distortion. The highest scored candidate is selected and added to the list of solution. The newly added connection serves as an additional anchor for a more accurate estimation of relative distortion. The connection candidates that are inconsistent with the selected connection are removed from the list. Subsequently, the rest of the candidates are re-evaluated again. Figure 1 presents the SenseAlign algorithm .

3. Example of running SenseAlign.

To illustrate how SenseAlign works, consider the pair of sentences (1e, 1c).

(1e) I caught a fish yesterday.
(1c) Zhuotian wuo budao yitiao yu.
yesterday I catch one fish.

Table 4 shows the connections that are considered in each iteration of the SenseAlign algorithm. Various factors used to evaluate connections are also given. Table 5 lists the connection in the final solution of alignment.

4. Experiments with SenseAlign

In this section, we present the experimental results of an implementation of SenseAlign and related algorithms. Approximately 25,000 bilingual example sentences from LecDOCE are used here as the training data. Here, the training data were used primarily to acquire rules by a greedy learner and to determine empirically probability functions of various factors. The algorithm's performance was then tested on the two sets of inside and outside data. The inside test consists of fifty sentence pairs from LecDOCE as input. The outside test are 416 sentence pairs from a book on English sentence patterns containing a comprehensive fifty-five sets of typical sentence patterns. However, the words in this outside test is somewhat more common, and, thereby, easier to align. This is evident from the slightly higher hit rate based on simple dictionary lookup.

The first experiment is designed to demonstrate the effectiveness of an naive algorithm (DictAlign) based on a bilingual dictionary. According to our results, although DictAlign produces high precision alignment, the coverage for both test sets is below 20%. However, if the thesaury effect is exploited, the coverage can be increased nearly three folds to about 40%, at the expense of a decrease around 10% in precision.

Table 1. Ten rules with the highest applicability

#	App.	Rule	Gloss for classes
1	642	Ma001, Hj63	moving / come, and go
2	459	Jh210, Di19	jobs, trade / work
3	440	Md108, Bo21	trains/car
4	418	Lg202, Eb28	new / new, fresh
5	367	Da003, Bn01	building, house/building
6	362	Gc060, Hi16	speaking / introduce
7	349	Fc050, Ed03	qualities / good, bad
8	310	Lh226, Tl18	measuring time / time
9	303	Ca002, Ab04	man and woman / baby
10	302	Fb020, Gb09	liking, loving / like, love

Table 2. Rules acquired from bilingual definitions for 12 senses of "bank" in LDOCE.

Sense & Definition	Rules
[1.n.1] land along the side of a river, lake, etc. 岸；堤	Ld099, Be03
[1.n.2] earth which is heaped up in a field or garden, often making a border or division. 田埂	Ld099, Bn12
[1.n.3] a mass of snow, clouds, mud, etc. 一堆； 一朝	Hb, Bb03
[1.n.4] a slope made at bends in a road or race-track, so that they are safer for cars to go round. 邊坡	Ld099, Be04
[1.n.5] = SANDBANK. 沙洲	Ld099, Be02
[2.v.1] (of a car or aircraft) to move with one side higher than the other, esp. when making a turn 傾斜轉彎	Nj295, Fd02
[3 n.1] a row, esp. of OARs in an ancient boat or KEYs on a TYPEWRITER. 一排	Hb, Dn08
[4.n.1] a place in which money is kept and paid out on demand, and where related activities go on. 銀行	Je104, Dm04
[4.n.2] (usu. in comb.) a place where something is held ready for use, esp. ORGANIC products of human origin for medical use. 儲存所	Je104, Bn17
[4.n.3] (a person who keeps) a supply of money or pieces for payment or use in a game of chance. 莊家	Je104, Dm04
[5.v.1] to put or keep (money) in a bank. 存於銀行	Je106, Hj40
[5.v.2] [esp. with] to keep one's money (esp. in the stated bank) 存款	Je106, Hj40

Table 3. Factor types with empirical probability

Factor	condition and probability			
FO	$f=1$	$f=2$	$f=3$	$f>3$
Prob	0.85	0.61	0.44	0.42
App	$A \geq .1$	$.1>A \geq .01$	$.01>A \geq .001$	$10^{-3}>A$
Prob	0.95	0.90	0.85	0.43
Spec	$S \geq 12$	$12>S \geq 11$	$11>S \geq 10$	$10>S$
Prob	0.95	0.85	0.77	0.35
R.D.	$rd=0$	$rd=1$	$rd=2$	$rd>2$
Prob	0.26	0.11	0.07	0.04
Sim	$Sim=1$	$1>Sim>.66$	$.66 \geq Sim \geq .2$	$Sim<.2$
Prob	0.94	0.42	0.35	0.12

Table 4. Various factors for connection candidates

Iteration	English Word	English POS	Chinese Word	Chinese POS	Rule		Fan-Out	*Sim*	*rd*	Spec	App.
1	yesterday	NR	昨天	Nd	Lh225	Tq23	1-1	1	4	11.2	0.0097
1	fish	NN	魚	Na	Ab032	Bi14	1-1	0.75	1	15.3	0.0017
1	I	PP	我	Nh	Gh280	Na02	1-1	1	1	0	0
1	I	PP	我	Nh	Gh280	Na05	1-1	1	1	0	0
1	fish	NN	魚	Na	Af100	Bi14	1-1	0.75	1	0	0
1	fish	NN	魚	Na	Ah120	Bi14	1-1	0.75	1	0	0
1	fish	NN	魚	Na	Ea017	Bi14	1-1	0.75	1	0	0
1	fish	NN	魚	Na	Eb031	Bi14	1-1	0.75	1	0	0
1	a	AT	一條	Nc	Nd098	Qa04	1-1	0.5	1	0	0
1	yesterday	NR	魚	Na	Lh225	Bi14	1-1	0	0	0	0
1	caught	VB	捕到	V+Di	Dc098	Hm05	1-1	0	1	0	0
1	fish	NN	昨天	Nd	Af100	Tq23	1-1	0	3	0	0
1	fish	NN	昨天	Nd	Ah120	Tq23	1-1	0	3	0	0
1	fish	NN	昨天	Nd	Ea017	Tq23	1-1	0	3	0	0
1	fish	NN	昨天	Nd	Eb031	Tq23	1-1	0	3	0	0
1	fish	NN	昨天	Nd	Ab032	Tq23	1-1	0	3	0	0
2	fish	NN	魚	Na	Ab032	Bi14	1-1	0.75	1	15.3	0.0017
2	I	PP	我	Nh	Gh280	Na02	1-1	1	1	0	0
2	I	PP	我	Nh	Gh280	Na05	1-1	1	1	0	0
2	fish	NN	魚	Na	Af100	Bi14	1-1	0.75	1	0	0
2	fish	NN	魚	Na	Ah120	Bi14	1-1	0.75	1	0	0
2	fish	NN	魚	Na	Ea017	Bi14	1-1	0.75	1	0	0
2	fish	NN	魚	Na	Eb031	Bi14	1-1	0.75	1	0	0
2	a	AT	一條	Nc	Nd098	Qa04	1-1	0.5	1	0	0
2	caught	VB	捕到	V+Di	Dc098	Hm05	1-1	0	1	0	0
3	I	PP	我	Nh	Gh280	Na02	1-1	1	0	0	0
3	I	PP	我	Nh	Gh280	Na05	1-1	1	0	0	0
3	a	AT	一條	Nc	Nd098	Qa04	1-1	0.5	0	0	0
3	caught	VB	捕到	V+Di	Dc098	Hm05	1-1	0	0	0	0
4	a	AT	一條	Nc	Nd098	Qa04	1-1	0.5	0	0	0
4	caught	VB	捕到	V+Di	Dc098	Hm05	1-1	0	0	0	0
5	caught	VB	捕到	V+Di	Dc098	Hm05	1-1	0	0	0	0

In our second experiment, we use SenseAlign described above for word alignment except that no bilingual dictionary is used. In our third experiment, we use the full SenseAlign to align the testing data. Table 6 indicates that acquired lexical information augmented and existing lexical information such as a bilingual dictionary can supplement each other to produce optimum alignment results. The generality of the approach is evident from the fact that the coverage and precision for the outside test are comparable with those of the inside test.

5. Discussions

5.1 Machine-readable lexical resources vs. corpora We believe the proposed algorithm addresses the problem of knowledge engineering bottleneck by using both corpora and machine readable lexical resources such as dictionaries and thesauri. The corpora provide us with training and testing materials, so that empirical knowledge can be derived and evaluated objectively. The thesauri provide classification that can be utilized to generalize the empirical knowledge gleaned from corpora.

SenseAlign achieves a degree of generality since a word pair can be accurately aligned, even when they occur rarely or only once in the corpus. This kind of generality is unattainable by statistically trained word-based models. Class-based models obviously offer advantages of smaller storage requirement and higher system efficiency. Such advantages do have their costs, for class-based models may be over-generalized and miss word-specific rules. However, work on class-based systems have indicated that the advantages outweigh the disadvantages.

5.2 Mutual information, and frequency. Gale and Church (1990) shows a near-miss example where ϕ^2, a χ^2-like statistic works better than mutual information for selecting strongly associated word pairs to use in word alignment. In their study, they contend that χ^2-like statistic works better because it uses co-

nonoccurrence and the number of sentences where one word occurs while the other does not which are often larger, more stable, and more indicative than co-occurrence used in mutual information.

The above-cited work's discussions of the χ^2-like statistic and the fan-in factor provide a valuable reference for this work. In our attempt to improve on low coverage of word-based approaches, we use simple filtering according to fan-out in the acquisition of class-based rules, in order to maximize both coverage and precision. The rules that provide the most instances of plausible connection is selected. This contrasts with approaches based on word-specific statistic where strongly associated word pairs selected may not have a strong presence in the data. This generally corresponds to the results from a recent work on a variety of tasks such as terminology extraction and structural disambiguation. Dallie, Gaussier and Lange (1994) demonstrated that simple criteria related to frequency coupled with a linguistic filter works better than mutual information for terminology extraction. Recent work involving structural disambiguation (Brill and Resnik 1994) also indicated that statistics related to frequency outperform mutual information and ϕ^2 statistic.

6. Concluding remarks

This paper has presented an algorithm capable of identifying words and their translation in a bilingual corpus. It is effective for specific linguistic reasons. The significant majority of words in bilingual sentences have diverging translation; those translations are not often found in a bilingual dictionary. However, those deviation are largely limited within the classes defined by thesauri. Therefore, by using a class-based approach, the problem's complexity can reduced in the sense that less number of candidates need to be considered with a greater likelihood of finding the correct translation.

In general, a slight amount of precision can apparently be expended to gain a substantial increase in applicability. Our results suggest that mixed strategies can yield a broad coverage and high precision word alignment and sense tagging system which can produce richer information for MT and NLP tasks such as word sense disambiguation. The word sense information can provide a certain degree of generality which is lacking in most statistical procedures. The algorithm's performance discussed here can definitely be improved by enhancing the various components of the algorithm, e.g., morphological analyses, bilingual dictionary, monolingual thesauri, and rule acquisition. However, this work has presented a workable core for processing bilingual corpus. The proposed algorithm can produce effective word-alignment results with

1. Read a pair of English-Chinese sentences.
2. Two dummies are replace to the left of the first and to the right of the last word of the source sentence. Similar two dummies are added to the target sentence. The left dummy in the source and target sentences align with each other. Similarly, the right dummies align with each other. This establishes anchor points for calculating the relative distortion score.
3. Perform the part-of-speech tagging and analysis for sentences in both languages.
4. Lookup the words in LEXICON and CILIN to determine the classes consistent with the part-of-speech analyses.
5. Follow the procedure in Section 2.3 to calculate a composite probability for each connection candidate according to fan-out, applicability, specificity of alignment rules, relative distortion, and dictionary evidence.
6. The highest scored candidate is selected and added to the list of alignment.
7. The connection candidates that are inconsistent with the selected connection are also removed from the candidate list.
8. The rest of the candidates are evaluated again according to the new list of connections.
9. The procedure iterates until all words in the source sentence are aligned.

Figure 1. Alignment Algorithm of SenseAlign

Table 5. The final alignment

English Word	English Code	Chinese Word	Chinese Code
I	Gh280	wuo	Na05
caught	De098	bu-dao	Hm05
a	Nd098	yi-tiao	Qa04
fish	Ab032	yu	Bi14
yesterday	Lh225	zuotian	Tq23

Table 6. Experimental Results

Inside Test				
No. Matched		# Hit	Coverage	Precision
DictAlign with *sim* = 1.0	59	56	15.3%	94.9%
DictAlign with *sim* > 0.67	113	100	29.4%	88.5%
DictAlign with *sim* > 0.5	151	124	39.2%	82.1%
SenseAlign without sim	237	213	61.7%	89.9%
Full SenseAlign	314	293	81.8%	93.3%
Outside Test				
No. Matched		# Hit	Coverage	Precision
DictAlign with *sim* = 1.0	499	486	16.8%	97.4%
DictAlign with *sim* > 0.67	970	865	32.7%	89.2%
DictAlign with *sim* > 0.5	1221	1046	41.1%	85.7%
SenseAlign without sim	1913	1721	66.8%	90.0%
Full SenseAlign	2424	2265	84.7%	93.4%

sense tagging which can provide a basis for such NLP tasks as word sense disambiguation (Chen and Chang 1994) and PP attachment (Chen and Chang 1995).

While this paper has specifically addressed only English-Chinese corpora, the linguistic issues that motivated the algorithm are quite general and are to a great degree language independent. If such a case is true, the algorithm presented here should be adaptable to other language pairs. The prospects for Japanese, in particular, seem highly promising. There are some work on alignment of English-Japanese texts using both dictionaries and statistics (Utsuro, Ikeda, Yamane, Matsumoto and Nagao 1994).

Acknowledgments

The authors would like to thank the National Science Concil of the Republic of China for financial support of this manuscript under Contract No. NSC 84-102-1211. Zebra Corporation and Longman Group are appreciated for the machine readable dictionary. Special thanks are due to Mathis H. C. Chen for work of preprocessing the MRD. Thanks are also due to Keh-Yih Su for many helpful comments on an early draft of this paper.

References

1. Brill, Eric and P. Resnik, (1994). A Rule based Approach to Prepositional Phrase Attachment, In *Proceedings of the 15th International Conference on Computational Linguistics*, 1198-1205, Kyoto Japan.
2. Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roosin, (1990). A Statistical Approach to Machine Translation, *Computational Linguistics*, 16:2, page 79-85.
3. Brown, P., S. Della Pietra, V. Della Pietra, and R. Mercer, (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, Vol. 19, No. 2, page 263-311.
4. Chang, J. S. and M. H. C. Chen, (1995). Structure Ambiguity and Conceptual Information Retrieval, In *Proceeding of Pacific Asia Conference on Language, Information and Computation*, page 16-23.
5. Chen, J. N. and J. S. Chang, (1994). Towards Generality and Modularity in Statistical Word Sense Disambiguation, In *Proceeding of Pacific Asia Conference on Formal and Computaional Linguistics*, page 45-48.
6. Dagan, Ido, K. W. Church and W. A. Gale, (1993). Robust Bilingual Word Alignment for Machine Aided Translation, In *Proceedings of the Workshop on Very Large Corpora : Academic and Industrial Perspectives*, page 1-8.
7. Daille, B., E. Gaussier and J.-M. Lange, (1994). Towards automatic extraction of monolingual and bilingual terminology, In *Proceedings of the International Conference on Computational Linguistics*, 515-521.
8. Fujii, Hideo and W. Bruce Croft, (1993). A Comparison of Indexing Techniques for Japanese Text Retrieval, In *Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 237-246.
9. Gale, W. and K. Church, (1993). A Program for Aligning Sentence in Bilingual Corpora, *Computational Linguistics*, 19(1), page 75-102.
10. Gale, W. A. and K. W. Church. (1991). Identifying Word Correspondences in Parallel Texts, In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, page 152-157, Pacific Grove, CA., February.
11. Gale, W. A., K. W. Church, and David Yarowsky, (1992). Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, 101-112, Montreal, Canada.Kay, Martin and Martin Roscheisen, (1993). Text-Translation Alignment, Computational Linguistics, Vol. 19, No. 1, page 121-142.
12. Longman, (1993). Longman English-Chinese Dictionary of Contemporary English, Published by Longman Group (Far East) Ltd., Hong Kong.
13. Matsumoto, Y. et al. (1993). Structural Matching of Parallel Texts, In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, page 1-30, Ohio, USA.
14. McArthur, T. (1992) Longman Lexicon of Contemporary English, Published by Longman Group (Far East) Ltd., Hong Kong.
15. Mei, J.J. et al., (1993). Tongyici Cilin (Word Forest of Synonyms), Tong Hua Publishing, Taipei, (traditional Chinese edition of a simplified Chinese edition published in 1984).
16. Proctor, Paul, (1988). Longman English-Chinese Dictionary of Contemporary English, Longman Group (Far East), Hong Kong.
17. Utsuro, T., H. Ikeda, M. Yamane, M. Matsumoto, and M. Nagao, (1994). Bilingual text matching using bilingual dictionary and statistics, In *Proceedings of the 15th International Conference on Computational Linguistics*, page 1076-1083, Kyoto, Japan.