Computational Linguistics and Chinese Language Processing
vol. 2, no. 2 August 1997, pp. 63-96.
© Computatinal Linguistics Soceity of R. O. C.

Aligning More Words with High Precision for Small Bilingual Corpora

Sue J. Ker* and Jason S. Chang+

Abstract

In this paper, we propose an algorithm for identifying each word with its translations in a sentence and translation pair. Previously proposed methods require enormous amounts of bilingual data to train statistical word-by-word translation models. By taking a word-based approach, these methods align frequent words with consistent translations at a high precision rate. However, less frequent words or words with diverse translations generally do not have statistically significant evidence for confident alignment. Consequently, incomplete or incorrect alignments occur. Here, we attempt to improve on the coverage using class-based rules. An automatic procedure for acquiring such rules is also described. Experimental results confirm that the algorithm can align over 85% of word pairs while maintaining a comparably high precision rate, even when a small corpus is used in training.

Keywords: Word alignment, machine readable dictionary and thesaurus, bilingual corpus, word sense disambiguation

1. Introduction

Research based on bilingual corpora has attracted an increasing amount of attention. Brown et al. (1990) advocated a statistical approach to machine translation (SMT) based on the bilingual Canadian Parliamentary debates. 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 *ST* and a *translation model* for mapping *ST* to a target text segment *TT*. They recommend using an *aligned* bilingual corpus to estimate the parameters in the translation model. Various levels of alignment resolution are possible; from section, paragraph, sentence, phrase, to word. In the process of word alignment, translation of each source word is identified. Their study focused primarily on identifying word-level alignment.

*Department of Computer Science, Soochow University, Taipei, Taiwan, ROC. E-mail:ksj@volans.cis.scu.edu.tw.
+Department of Computer Science, National Tsing Hua University, Hsin-chu, Taiwan, ROC.

In the context of SMT, Brown et al. (1993) presented a series of five models for estimating translation probability. The first two models have been used in research on word alignment. Model 1 assumes that translation probability depends only on *lexical translation probability*. Model 2 enhances Model 1 by considering the dependence of translation probability on the *distortion probability*.

There are statistical tools that can help determine the relative association strength of bilingual word pairs with respect to translatability. Gale and Church (1991) used ϕ^2 to identify the word correspondence from a bilingual corpus while Fung and Church (1994) proposed a K-vec approach, which is based on a k-way partitioning of the bilingual corpus, to acquire a bilingual lexicon. Such tools usually provide low coverage due to the fact that low frequency words are in the majority and high frequency words tend to have diverse translations.

Estimates of word-to-word translation probability based on lexical co-occurrence (Gale and Church, 1991; Kay and Röscheisen, 1993; Fung and Church, 1994; Fung and McKeown, 1994; Utsuro, Ikeda, Yamane, Matsumoto and Nagao, 1994; Smadja, McKeown and Hatzivassiloglou, 1996) are highly unreliable for sparse data. In general, some kind of filtering is required to reduce noise. This leads to a low coverage rate. For instance, Gale and Church (1991) reported that their ϕ^2 method produced highly precise (95%) alignment for only 61% of the words in 800 sentences tested. Wu and Xia (1994) employed the EM algorithm (Dempster, Laird and Rubin, 1977) to find the optimal word alignment from a sentence-aligned corpus. The authors claim that they obtained a high precision rate of between 86% and 96%. However, the coverage rate was not reported.

The above survey clearly indicates that word-based methods offer limited lexical coverage even after they are trained with a very large bilingual corpus. For most applications, low coverage is just as serious a problem as low precision. For aligned corpora to be useful for NLP tasks, such as machine translation and word sense disambiguation, a coverage rate higher than 60% is desirable, even at the expense of a slightly lower precision rate. A bilingual corpus with all instances of polysemous words correctly connected to their translations provides valuable training material for developing a WSD system (Gale, Church, and Yarowsky, 1992).

In this paper, we propose a word-alignment algorithm, *SenseAlign*, based on classes derived from sense-related categories in existing thesauri. *SenseAlign* relies on an automatic procedure to acquire class-based alignment rules (Ker and Chang 1996). To make even broader coverage possible, we exploit additional sources of knowledge; connections that are evident from some, but not necessarily all, knowledge sources can

still be aligned. The algorithm aligns over 85% of word pairs with a comparably high precision rate of 90%.

The rest of this paper is organized as follows. The next section describes *SenseAlign* and discusses its main components. Section 3 provides illustrative examples taken from the Longman English-Chinese Dictionary of Contemporary English (Longman 1992, LecDOCE, henceforth). Section 4 summarizes the experimental results. Additionally, typological and quantitative error analyses are also reported. Section 5 compares *SenseAlign* to several other approaches that have been proposed in the literature of computational linguistics. Finally, Section 6 considers ways in which the proposed algorithm might be extended and improved.

2. The Word Alignment Algorithm

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

2.1 The Thesauri

In this work, the categories for Chinese text are taken from a thesaurus for Mandarin Chinese (Mei, Zhu, Gao, and Yin 1993, CILIN henceforth). The categories for English text are taken from the Longman Lexicon of Contemporary English (McArthur 1992, LLOCE henceforth). The division of words into semantic categories in the two thesauri is somewhat different. The categories in CILIN are organized as a conceptual ontology of three levels: gross categories, intermediate categories and detailed categories.

Unlike CILIN, the categories of LLOCE are organized primarily according to subject matter. The LLOCE categories are also organized as three levels: subjects, titles and sets. In the first level, fourteen major subjects are denoted with reference letters from *A* to *N*. For detailed descriptions of CILIN and LLOCE see Appendices A and B.

2.2 Lexical Analyses

Two taggers are utilized to resolve part-of-speech ambiguity. Morphological and idiom analyses are also performed to determine the lexical unit and lexeme. Only thesaurus categories consistent with the part-of-speech determined in the analysis are considered in subsequent processes.

The part-of-speech taggers for the two languages involved are built using a strategy proposed by Brill (1992). We use the tag set in the Brown Corpus for the English tagger

and the part-of-speech system proposed by Chao (1968) for the Chinese tagger. Tables 1.1 and 1.2 present the two tag sets.

To eliminate the difficult cases of 0-1 *fertility* (one target word aligned with nothing in the source sentence), certain morpho-syntactical constructs in Chinese are identified. They are, mainly, Chinese constructions (see Table 2) that have no parallel in English, such as direction or phrase complements (Di, VH, or Ng) following a verb and measure nouns (Nf) following a determinant/quantifier (Ne). Tables 3.1 and 3.2 list the outputs of the two taggers for Examples (1e, 1c).

POS	Meaning	POS	Meaning
AB	Pre-qualifier	NP	Proper noun
AT	Article	NR	Adverbial noun
BE	Be	PN	Nominal Pronoun
CC	Coordinating conj.	PP	Personal Pronoun
CD	Cardinal numeral	RB	Adverb
CS	Subordinate conj.	RN	Nominal adverb
DT	Determiner	RP	Adverb or particle
IN	Preposition	TO	Infinite marker
JJ	Adjective	UH	Interjection
NN	Noun	VB	Verb

Table 1.1 *The parts-of-speech used in the English*

POS	Meaning	POS	Meaning
A	Adjective	Na	Common noun
C	Conjunction	Nb	Proper noun (person)
Da	Quantity adverb	Nc	Proper noun (locative)
Db	Judgment adverb	Nd	Temporal
Dc	Negation adverb	Ne	Determinant or quantifier
Dd	Temporal adverb	Nf	Measure noun
De	Degree adverb	Ng	Locative noun
Df	Locative adverb	P	Preposition
Di	Aspect	V	Verb
Dj	Question adverb	VH	State Verb

Table 1.2 *The parts-of-speech used in the Chinese tagger*

Constructs	Example
Verb+Aspect	捉了(zhuo-le, caught)
Verb+Direction	捉來(zhuo-lai, caught and bring in)
Verb+State	捉完(zhuo-uan, finished catching)
Determinant+Measure	這個(zhe-ge, this)
Quantifier+Measure	一個(yi-ge, one piece)

Table 2 *Verb-complement and determinant-measure constructs*

English Word	POS	Reference Code
lightning	NN	Lc058
usually	RB	Nc056
accompanies	VB	Kb032
thunder	NN	Lc058

Table 3.1 *Tagging results of Example (1e)*

Chinese Word	POS	Reference Code
雷聲 (leisheng)	Na	Bf06
通常 (tongchang)	Dd	Ka10
隨著 (suizhe)	V+Di	Hj36
閃電 (shandian)	Na	Bf06
而 (er)	C	Kc02, Kc03, Kc08
來 (lai)	V	Hj12, Hj63, Jd07

Table 3.2 *Tagging results of Example (1c)*

(1e) Lightening usually accompanies thunder.

(1c)	雷聲	通常	隨著	閃電	而	來。
	leisheng	tongchang	suizhe	shandian	er	lai
	thunder	usually	accompany	lightning	and	come

2.3 The Greedy Learner

The main mechanism of *SenseAlign* is the class-based alignment rules. Those rules form a subset of the Cartesian product of the categories in the two thesauri. We were inspired by the revision model proposed by Brill and Resnik (1994) in designing an automatic acquisition procedure for alignment rules. The procedure employs the greedy method to find a set of rules capable of providing optimal alignment in a bilingual corpus.

The rule capable of providing the most instances of plausible alignment is preferred and selected first. First, the bilingual example sentences go through some lexical analyses. The lexemes are then looked up in the thesauri to find the possible categories under which they may be listed. At this stage, no information is available regarding what classes of words are likely to align with each other. Second, we randomly match up words and their categories across the sentence pairs to form tentative *alignment rules*. Third, after producing tentative alignment rules for all the sentences, we make a conservative estimate of *applicability*. The rule with the highest *estimated applicability* is selected. Sentences where the rule applies are identified. The matched connections (*s*, *t*) in those sentences are removed. In addition, connections (*s*, *t'*) and (*s'*, *t*) for all $s' \neq s$ and all $t' \neq t$ are removed because they are inconsistent with the selection of (*s*, *t*). The acquisition process is repeated for the remaining data until applicability of the best rule runs below a certain threshold. The learning algorithm can be applied to acquire rules having different levels of resolution.We have run the learning algorithm on the 25,000 bilingual examples in LecDOCE. This procedure for learning rules has been applied to the detailed categories of CILIN and the topical sets of LLOCE to produce 392 rules. Table 4.1 and Table 4.2 present the ten rules with high and middle applicability. Figure 1 shows the accumulative applicability distributions of 392 rules. Obviously, these 392 rules do not cover all English words, nor all Chinese words. To remedy this problem, the procedure is repeated for broader 2-letter classes represented by topics in LLOCE and intermediate categories in CILIN. See Table 4.3 for three rules acquired on the 2-letter level. Table 5 presents the number of rules acquired on two levels of resolution.

Rule#	#App.	POS		Rule	Gloss for LLOCE	Gloss for CILIN
1	642	VB	V	Ma001, Hj63	moving, coming, and going	來(lai, come), 去(qu, go)
2	459	NN	Na	Jh210, Di19	jobs, trade and professions	職業(zhie, job)
3	440	NN	Na	Md108, Bo21	trains	車(che, car)
4	418	JJ	A	Lg202, Eb28	new	新(xin, new), 新鮮(xinxian, fresh)
5	367	NN	Na	Da003, Bn01	things built and lived in	建築(jianzhu, building)
6	362	VB	V	Gc060, Hi16	speaking, and telling	介紹(jieshao, introduce)
7	349	JJ	A	Fc050, Ed03	the right qualities	好(hao, good), 壞 (huai, bad)
8	310	NN	Na	Lh226, Tl18	measuring time	年(nian, year)
9	303	NN	Na	Ca002, Ab04	man and woman	嬰兒(ienger, baby)
10	302	VB	V	Fb020, Gb09	liking and loving	喜歡(xihuan, like), 愛(ai, love)

Table 4.1 *Ten rules with high applicability*

Rule#	#App.	POS		Rule	Gloss for LLOCE	Gloss for CILIN
101	95	NN	Na	Ab030, Ba02	living things	生物 (shengwu, living things)
102	95	JJ	A	Jc063, Ed26	relating to measurement	尊貴 (zungui, nobility)
103	94	JJ	A	Bh110, Ed03	good bodily condition	好 (hao, good), 壞 (huai, bad)
104	93	VB	V	Ma004, Id21	leaving and setting out	靠近 (kaojin, coming on)
105	91	NN	Na	Mh202, Bc02	front, back and sides	邊 (bian, side) 角 (jiao, corner)
106	91	JJ	A	Nd096, Ua01	much, many	非常 (feichang, very)
107	90	NN	Na	Kh196, Bd03	cricket	地球 (diqiu, earth)
108	90	NN	Na	Cn270, Di11	war and peace	戰爭 (zhanzheng, war)
109	90	NN	Na	Lh226, Tp22	measuring time	星期 (xingqing, week)
110	89	NN	Na	Gf233, Dd15	word and names	名稱 (mingcheng, appellation)

Table 4.2 *Ten rules with the middle applicability*

Rule#	#App.	POS		Rule	Gloss for LLOCE	Gloss for CILIN
1	1628	VB	V	Ma, Hj	moving, coming, and going	生活(shenghuo, activities in daily life)
2	1251	VB	V	Gc, Hi	Communicating	社交(shejiao, social activities)
3	980	JJ	A	Mh, Ed	locating and direction	性質(xingzh, property)

Table 4.3 *Three rules with the high estimated applicability on the 2-letter level*

Class level	No. of rules acquired
set vs detailed category	392
Title vs intermediate category	3

Table 5 *The number of selected pairs*

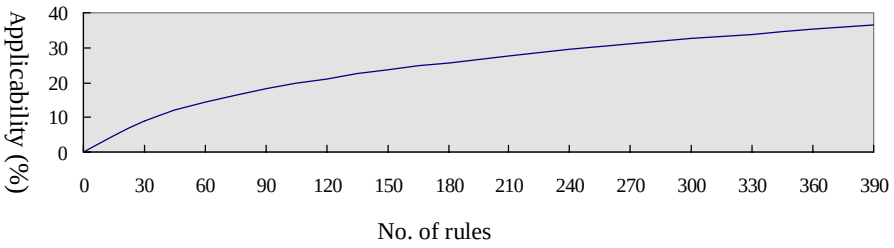

Figure 1 *The cumulative applicability of the acquired rules.*

2.4 Fan-out

When matching words in *ST* and *TT* against a rule, we use the term *fan-out* to denote the number of words that match the rule. For instance, for a rule $r = (C, D)$, and a sentence pair (*ST*, *TT*), the rule *r* has a fan-out of *n-m* if there are *n* and *m* words in *ST* and *TT* listed under classes *C* and *D*, respectively. The *degree* of fan-out of a connection applying rule *r* is given by

$$F = n \times m.$$

For instance, the learner produces the connections shown in Table 6 for Example 1. Both "lightning" and "thunder" in (1e) are listed under LLOCE set *Lc058* (thunder and lightning) while both "雷聲 (leisheng, thunder-sound)" and "閃電 (shandian, flash-electricity)" in (1c) are listed under CILIN category *Bf06* (雷 (lei, thunder); 閃電 (shandian, flash-electricity)). Therefore, the rule (*Lc058*, *Bf06*) applies to the first four connections shown in Table 6. The rule is said to have a fan-out value of 2-2 in sentence pair (1e, 1c). On the other hand, the rule (*Nc056*, *Ka10*) applies to only one connection (*usually*, 通常 (usually, normally)) with a 1-1 fan-out.

English POS	English Code	English Word	Chinese POS	Chinese Code	Chinese Word	Rule	Fan-out
NN	Lc058	lightning	Na	Bf06	雷聲(leisheng, thunder)	(Lc058, Bf06)	2-2
NN	Lc058	lightning	Na	Bf06	閃電(shandian, lightning)	(Lc058, Bf06)	2-2
NN	Lc058	thunder	Na	Bf06	雷聲(leisheng, thunder)	(Lc058, Bf06)	2-2
NN	Lc058	thunder	Na	Bf06	閃電(shandian, lightning)	(Lc058, Bf06)	2-2
RB	Nc056	usually	Dd	Ka10	通常(tongchang, usually)	(Nc056, Ka10)	1-1
VB	Kb032	accompany	V+Di	Hj36	隨著(suizhe, accompany)	(Kb032, Hj36)	1-1
VB	Kb032	accompany	V	Hj12	來 (lai, come)	(Kb032, Hj12)	1-1
VB	Kb032	accompany	V	Hj63	來 (lai come)	(Kb032, Hj63)	1-1
VB	Kb032	accompany	V	Jd07	來 (lai come)	(Kb032, Jd07)	1-1
VB	Mb053	accompany	V+Di	Hj36	隨著(suizhe, accompany)	(Mb053, Hj36)	1-1
VB	Mb053	accompany	V	Hj12	來 (lai, come)	(Mb053, Hj12)	1-1
VB	Mb053	accompany	V	Hj63	來 (lai, come)	(Mb053, Hj63)	1-1
VB	Mb053	accompany	V	Jd07	來 (lai, come)	(Mb053, Jd07)	1-1

Talbe 6 *The tentative connections for Example (1e, 1c)*

2.5 Specificity

Some rules are more *specific* because they apply to two small sets of words in the thesauri. The more specific a rule, the more likely it applies to words that are interchangeable translations. Therefore, we define the *specificity S* for a connection (*s*, *t*) to which a rule $r = (C, D)$ is applicable as follows:

$$S = -\log(\Pr(x \in C) \times \Pr(y \in D)) \quad \text{if } (s, t) \in (C, D) \in R,$$
$$= 0 \quad \text{otherwise,}$$

where R is the set of acquired rules; $\Pr(x \in C)$ and $\Pr(y \in D)$ are the probabilities of generating words x and y in classes C and D, respectively. Thus, the specificity S of r reflects the probability of generating, by chance, a pair of words (s, t), $s \in C$ and $t \in D$ to which r is applicable.

Assuming that the distribution of the words in a given class is uniform, the *degree* of specificity of a connection applying a rule r is

$$S_r = -\log\left(\frac{1}{W_e \times W_e}\right), \text{ where } r = (e, c),$$

W_e = the number of English words in class e,
W_c = the number of Chinese words in class c.

For instance, consider Example (2e, 2c), where the rule (*Ac053, Bi08*) is used to connect "*cat*" to " 貓 ". The number of words in LLOCE class "Ac053" is 34, and the number of words in CILIN class "Bi08" is 42. Therefore, we obtain the following specificity:

$$S_{(\text{Ac053, Bi08})} = -\log\left(\frac{1}{34 \times 42}\right) = 10.48.$$

(2e) I only knew that it is the dog not the cat that bit me.

(2c)	我	只	知道	咬	我	的	是	狗	不是	貓。
	wuo	zhi	zhidao	yiao	wuo	de	shi	gou	bushi	mao.
	I	only	know	bit	I	DE	is	dog	NOT	cat.

2.6 Applicability

The acquired rules may have different degrees of *applicability*. Applicability refers to the number of instances of word pairs in the bilingual corpus to which is applicable. The higher the applicability an alignment rule has, the more reliable are the connections it predicts. Furthermore, including the factor of applicability also results in more connections being chosen. Therefore, we define *applicability* for a connection (*s*, *t*) to which a rule $r = (C, D)$ is applicable as follows:

$$\text{Applicability}: A_r = \frac{C_r}{B},$$

where C_r is the number of connections for which the rule *r* is applicable in the corpus, and *B* is the number of bilingual sentences in the corpus.

For instance, consider the rule (*Ac053, Bi08*) in Example (2e, 2c) again. There are 55 instances of connections in a corpus of 25,000 sentences to which (*Ac053, Bi08*) is applicable. Therefore, we can obtain the following applicability:

$$A_{(\text{Ac053, Bi08})} = \frac{55}{25{,}000} = 0.0022.$$

2.7 Relative Distortion

We use a new distortion measure in addition to the alignment rules to evaluate the plausibility of a connection candidate. To be more specific, we adopt the relative distortion to form a model of position which is much smaller than that obtained based on absolute position. This choice is based on the observation that many language con-

structions are preserved in the translation process. Therefore, the target position of a connection relative to that of some connection in the same construction has a much smaller variance in statistical distribution. However, we use an approximation of relative distortion for lack of structural analysis. Assuming that some connections have been selected, we can always evaluate a candidate (*s*, *t*) relative to these connections. Let *s* and *t* be the *i*th and *j*th words in *ST* and *TT*, respectively. There exist two closest connections, (i_L, j_L) and (i_R, j_R), on both sides of *s*. Relative distortion **rd**(*s*, *t*) is approximated using the following formula:

$$\mathbf{rd}(s, t) = \min(|d_L|, |d_R|) \quad \text{where } d_L = (j - j_L) - (i - i_L),$$
$$d_R = (i - i_R) - (j - j_R).$$

Empirical data confirm that connections with small *rd* values are more likely to be correct. Figure 2 indicates that a candidate with 0 *rd* values is much more probable (nearly .80) as a correct connection than is a candidate with 0 absolute distortion (.43).

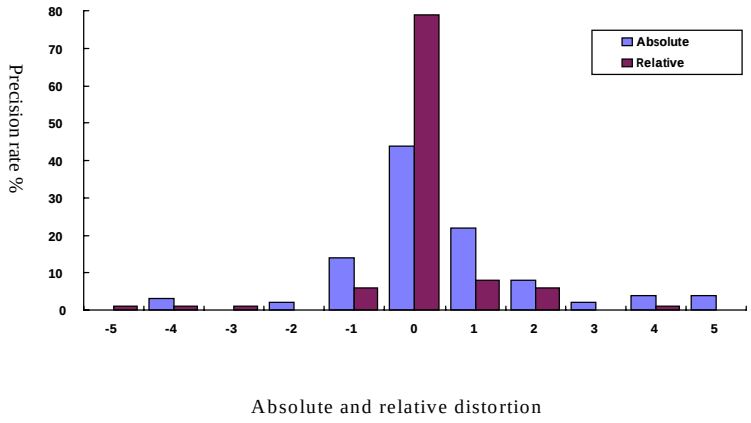

Figure 2 *Precision rates for candidates with different values of distortion*

In the following, Example (3e, 3c) demonstrates how distortions play an influential role in determining the correct connection.

(3e) Please$_1$ answer$_2$ all$_3$ **questions**$_4$ on$_5$ this$_6$ **list**$_7$.

(3c)

請 $_1$	回答 $_2$	本 $_3$	表 $_4$	上 $_5$	之 $_6$	所有 $_7$	問題 $_8$
qing	huida	ben	**biao**	shang	zh	suoiou	**wunti**
Please	answer	this	**list**	on	CTM	all	**question**

After lexical processing for both languages is completed, the initial estimations of relative distortion for all possible connection candidates are calculated (Table 7.1). Notably, many true connections receive a value of 3 for their relative distortion. These large *rd* values are due to forward transfer of the prepositional phrase "*on this list*" to the front of the attached noun phrase, "all questions." However, the *rd* estimates become more and more accurate with each iteration. For instance, if the connection (*question*, 問題 *wunti*) is selected, the *rd* for the candidate (all, 所有 suoyiou) is re-evaluated correctly at 0. Similarly, if (*list*, 表 *biao*) is selected, the *rd* for the candidate (*this*, 本 *ben*) is also re-evaluated at 0. Tables 7.1 through 7.3 provide further details.

English	Position	POS	Chinese	Position	POS	d_L	d_R	*rd*
answer	2	VB	回答 (huida, answer)	2	V	0	1	0
all	3	AT	所有 (suoyiou, all)	7	Ne	4	-3	3
question	4	NN	表 (biao, list)	4	Na	0	1	0
question	**4**	**NN**	**問題 (wunti, question)**	**8**	**Na**	**4**	**-3**	**3**
on	5	IN	上 (shang, up)	5	Ng	0	1	0
this	6	AT	本 (ben, this)	3	Ne	-3	4	3
list	**7**	**NN**	**表 (biao, list)**	**4**	**Na**	**-3**	**4**	**3**
list	7	NN	問題 (wunti, question)	8	Na	1	0	0

Table 7.1 *The relative distortion of word pairs (Initially)*

English	Position	POS	Chinese	Position	POS	d_L	d_R	*rd*
answer	2	VB	回答 (huida, answer)	2	V	0	4	0
all	**3**	**AT**	**所有 (suoyiou, all)**	**7**	**Ne**	**4**	**0**	**0**
on	5	IN	上 (shang, up)	5	Ng	-4	1	1
this	6	AT	本 (ben, this)	3	Ne	-7	4	4
list	7	NN	表 (biao, list)	4	Na	-7	4	4

Table 7.2 *The relative distortion after selection of (question,* 問題）

English	Position	POS	Chinese	Position	POS	d_L	d_R	*rd*
answer	2	VB	回答 (huida, answer)	2	V	0	-3	0
all	3	AT	所有 (suoyiou, all)	7	Ne	4	-7	4
question	**4**	**NN**	**問題 (wunti, question)**	**8**	**Na**	**4**	**-7**	**4**
on	**5**	**IN**	**上 (shang, up)**	**5**	**Ng**	**0**	**-3**	**0**
this	**6**	**AT**	**本 (ben, this)**	**3**	**Ne**	**-3**	**0**	**0**

Table 7.3 *The relative distortion if (list,* 表 *) is selected initially.*

2.8 Similarity between Connection Target and Dictionary Translations

Dictionary translations are an important knowledge source for word alignment. Approximately 40% of the targets in correct connections have at least one Chinese character in common with dictionary translations for the corresponding source word (Ker and Chang, 1997). Such a target and translation can be thought of as synonyms. Consider Example (1e, 1c) again. Four translations are listed in LecDOCE for "*accompany*": 1. " 伴 " (*ban*, keep somebody company), 2. " 陪 " (*pei*, to be with somebody), 3. " 隨 " (*suei*, to follow), and 4. " 伴奏 " (*banzou*, to make supporting music for). The connection target " 隨著 " (*suizhe*, to follow + ASP) of "*accompany*" has one character in common with the third translation.

To exploit this thesaury effect (Fujii and Croft 1993) in translation, we need a way to measure the similarity between words. The similarity measure of the Dice coefficient (Dice 1945) seems to be a good choice. Equation (1) shows the formulation of the Dice coefficient. An unweighted version of the Dice coefficient shown in Equation (2) can also be used for simplicity:

$$(1)\quad \frac{2\sum_{k=1}^{|E|} w(E_k)}{\sum_{k=1}^{|C|} w(C_k) + \sum_{k=1}^{|D|} w(D_k)} \quad ,$$

$$(2)\quad \frac{2\,|E|}{|C| + |D|} \quad ,$$

where $|C|$ = the length of *C*, Mandarin translation in connection,

$|D|$ = the length of *D*, Mandarin morphemes in dictionary,

$|E|$ = the length of E, common Mandarin morpheme in *C* and *D*,

$w(C_k)$ = weight of the k-th Mandarin morpheme in *C*,

$w(D_k)$ = weight of the k-th Mandarin morpheme in *D*,

$w(E_k)$ = weight of the k-th Mandarin morpheme in *E*.

In Example (4e, 4c), connection candidates such as (*yesterday*, 昨天 *zhuotian*), (*today*, 今天 *jintian*), and (*feel*, 覺得 *juede*) match the entries in LecDOCE completely; therefore, they receive a similarity value of 1. On the other hand, the connection (*ill*, 不舒服 (*bushufu*, not comfortable)) receives a similarity score of 0.33 since it shares the Chinese character " 不 " (*bu*, no) with an LecDOCE translation of "ill," " 不好的 " (*buhaode*, not good).

(4e) Yesterday I was ill but today I am feeling A-1.

(4c) 昨天	我	不舒服	今天	卻	覺得	很	好。
zhuotian	wuo	bushufu	jintian	que	juede	hen	hao
yesterday	I	not comfortable	today	but	feel	very	good

2.9 Evaluation of Connection Candidates

As mentioned earlier, Brown et al's Model 2 (1993) stipulates that a connection be given a probability value as the product of lexical translation probability and distortion probability under the assumption of independency. In the same spirit, we give a composite probabilistic value for each connection candidate by multiplying the probabilities of these factors. Therefore, the formula of evaluating the composite probability is as follows:

$$\text{Prob}(s, t) = \max_{\substack{s\in c \\ t\in d}} \text{Prob}(s, t|\text{fan\_out}(c,d)) \times \text{Prob}(s,t|\text{rd}(s,t)) \times \text{Prob}(s,t|A_r(c,d)) \times \text{Prob}(s,t|S_r(c,d)) \times \text{Prob}(s,t|sim(s,t))$$

where s = source word,

t = target word,

c = an LLOCE class of which s is a member, and

d = a CILIN class of which t is a member.

The probabilities of these factors are estimated according to the principle of maximum likelihood estimation (MLE). For instance, if there are *k* connections in a sample of *n* candidates (*s, t*) whose degree of fan-out is *f*, then the alignment probability Prob(*s, t* | *f*) for each (*s, t*) is given the same MLE value, i.e. Prob(*s, t* | *f*) = *k/n* for all pair (*s, t*). By using a small sample of a few hundred sentences, the MLE probabilities for various factors can be estimated quite reliably. Table 8 summarizes the MLE probabilistic values

obtained using 200 manually aligned sentences from the LecDOCE.

Factor type	Condition and empirically estimated probability				
Fan-out	condition	$f = 1$	$f = 2$	$f = 3$	$f > 3$
	probability	0.85	0.61	0.44	0.42
Applicability	condition	$A \geq 10^{-2}$	$10^{-2} > A \geq 10^{-3}$	$10^{-3} > A \geq 10^{-4}$	$10^{-4} > A$
	probability	0.95	0.90	0.85	0.43
Specificity	condition	$10 > S > 0$	$12 > S \geq 10$	$S \geq 12$	$S = 0$
	probability	0.95	0.77	0.45	0.20
Relative distortion	condition	$rd = 0$	$rd = 1$	$rd = 2$	$rd > 2$
	probability	0.26	0.11	0.07	0.04
Similarity to Dictionary translation	condition	$Sim = 1.0$	$1.0 > Sim \geq 0.66$	$0.66 > Sim \geq 0.2$	$Sim < 0.2$
	probability	0.94	0.42	0.35	0.12

Table 8 *Factor types with MLE probability.*

For instance, consider the Example (5e, 5c), focusing on the word pair (yesterday, 昨天) :

(5e) I caught a fish yesterday.

(5c) 昨天 我 捕到 一條 魚。
zhuotian wuo budao yitiao yu.
yesterday I catch one fish.

Prob(*yesterday*, 昨天)= Max (Prob(*yesterday*, 昨天 |fan-out(*Lh*225,*Tq*23))
×Prob(yesterday, 昨天 |rd(yesterday, 昨天))
× prob(yesterday, 昨天 |Ar(*Lh*225,*Tq*23))
× Prob(yesterday, 昨天 |Sr(*Lh*225,*Tq*23))
×Prob(yesterday, 昨天 |sim(yesterday, 昨天)))
= Max (Prob(*yesterday*, 昨天 |f=1)
× Prob(yesterday, 昨天 |rd =4)
× prob(yesterday, 昨天 |A_r = 0.0097)
×Prob(yesterday, 昨天 |S_r = 11.2)

$\times$Prob(yesterday, 昨天 |*sim* = 1.0))

= Max (0.85 $\times$ 0.04$\times$ 0.90$\times$ 0.77 $\times$0.94)
= 0.0022

Distortion is used in our algorithm in a way similar to that in Gale and Church (1991). However, our consideration of the right neighbor of a candidate in addition to the left one realizes a much tighter approximation of relative distortion. Other factors are also introduced to distinguish the case in which fan-out and distortion alone can not determine the right alignment. Empirical data have indicated that connections suggested by using a rule with a higher degree of *specificity* or *applicability* are more likely to be correct. The preference for using rules with higher applicability also has the effect of boosting the overall hit rate. Applicability and specificity are analogous to term frequency and inverse document frequency, the two weighting factors most widely used in IR research.

2.10 Alignment Algorithm

Our algorithm for word alignment is a decision procedure for selecting the preferred connection from a list of candidates. Initial anchors for calculating relative distortion can be established by placing two dummies at the front and end of *ST* and *TT*. The initial list contains two connections that are formed from those four dummies. Consequently, the highest scoring candidate is selected and added to the list of solutions. The newly added connection serves as an additional anchor for more accurate estimation of relative distortion. The connection candidates that are inconsistent with the selected connection are removed from the list. Consequently, the rest of the candidates are re-evaluated again. Figure 3 presents the *SenseAlign* algorithm. Table 9 summarizes all of the factors used in *SenseAlign*.

Fan-out: $f = n \times m$ where n = the number of C-class words in *ST*,
m= the number of D-class words in *TT*.

Specificity: $S_r = -\log(\frac{1}{W_e \times W_c})$, where $r = (e, c)$,
W_e = the number of English words in class e,
W_c = the uumber of Chinese words in class c.

Applicability: $A_r = \frac{C_r}{B}$, where C_r = the number of connections for which r is applicable in the corpus,
B = the number of bilingual sentences in the corpus.

Relative Distortion:

$\mathbf{rd}(i, j) = \min(|d_L|, |d_R|)$,
$d_L = (j - j_L) - (i - i_L)$, where i = the subscript of the source sentence,
$d_R = (i - i_R) - (j - j_R)$, j = the subscript of the target sentence,
i_L = the position of the closest connection to the left of the i-th word in the source sentence,
j_L = the aligned target word of i_L,
i_R = the position of the closest connection to the right of the i-th word in the source sentence,
j_R = the aligned target word of i_R.

Similarity between the connection target and dictionary translation:

$sim = \frac{2|E|}{|C| + |D|}$, where $|C|$ = the number of morphemes in C,
$|D|$ = the number of morphemes in D,
$|E|$ = the number of common morphemes in C and D.

Table 9 *Summary of factors and formula used in SenseAlign Algorithm*

1. Read a pair of English-Chinese sentences.
2. Place two dummies to the left of the first and to the right of the last word of the source sentence. Two similar 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 the sentence in both languages.
4. Lookup the words in LLOCE and CILIN to determine the classes consistent with the part-of-speech analyses.
5. Follow the procedure in Section 2.9 to calculate a composite probability for each connection candidate according to fan-out, applicability, specificity of alignment rules, relative distortion, and dictionary evidence.
6. Select the highest scoring candidate and add it to the alignment list.
7. Remove the connection candidates that are inconsistent with the selected connection from the candidate list.
8. Re-evaluate the rest of the candidates again according to the new list of connections.
9. Repeat Steps 4-8 until all words in the source sentence are aligned or every remaining word pair is associated with a score lower than some preset threshold *h*.

Figure 3
Alignment algorithm for SenseAlign

3. Illustrative Examples

To illustrate how *SenseAlign* works, consider the sentence pair (5e, 5c) mentioned previously:

(5e) I caught a fish yesterday.

(5c) 昨天 我 捕到 一條 魚。
zhuotian wuo budao yitiao yu.
yesterday I catch one fish.

After Step 3 of *SenseAlign* is executed, the algorithm produces the analyses shown in

Table 10.1 and Table 10.2. Table 11 provides the glossary of class codes involved. Consequently, the algorithm selects the highest-scored connection, (*yesterday,* 昨天 *zhuotian*). Next, this connection and other inconsistent connections are removed. In the subsequent iterations, the connections (fish, 魚 yu), (I, 我 wuo) and (a, 一條 yi-tiao) are selected. Table 12 shows the remaining connections after each iteration. Table 13 summarizes the connections selected to form a word alignment solution.

English word	POS	Class code (e)	W_e
I	PP	Gh280	13
caught	V	De098	12
a	AT	Nd098	6
fish	NN	Af100, Ah120, Ab032, Ea017, Eb031	32, 6, 22, 21, 9
Yesterday	NR	Lh225	8

Table 10.1 Results of lexical processing for example sentence 5e.

Chinese Word	POS	Class code (c)	W_c
昨天	Nd	Tq23	47
我	Nh	Na02, Na05	76, 21
捕到	V+Di	Hm05	27
一條	Ne	Qa04	52
魚	Na	Bi14	15

Table 10.2 *Results of lexical processing for example sentence 5c.*

Code	Gloss	Code	Gloss
Gh280	personal Pronouns	Qa04	number
Nd098	some and any	Bi14	fish, shrimp
Ab032	kinds of living creature	Tq23	today, yesterday, tomorrow
Lh225	time	Na02	I, we
De098	taking and catching things	Na05	oneself, others, somebody
Af100	common fish	Hm05	arrest, release
Ah120	parts around the head and neck		
Ea017	courses in meals		
Eb031	meat, etc.		

Table 11 *Glossary of class codes relevant to the Example (5e, 5c).*

English POS	English Word	English Code	Chinese POS	Chinese Code	Chinese Word	Fan-Out	*Sim*	*rd*	S	A	Prob
After initial alignment of dummies											
NR	yesterday	Lh225	Nd	Tq23	昨天	1-1	1	4	11.2	0.0097	.0221
NN	fish	Ab032	Na	Bi14	魚	1-1	0.75	1	15.3	0.0017	.0159
PP	I	Gh280	Nh	Na02	我	1-1	1	1	0	0	.0076
PP	I	Gh280	Nh	Na05	我	1-1	1	1	0	0	.0076
NN	fish	Af100	Na	Bi14	魚	1-1	0.75	1	0	0	.0034
NN	fish	Ah120	Na	Bi14	魚	1-1	0.75	1	0	0	.0034
NN	fish	Ea017	Na	Bi14	魚	1-1	0.75	1	0	0	.0034
NN	fish	Eb031	Na	Bi14	魚	1-1	0.75	1	0	0	.0034
AT	a	Nd098	Ne	Qa04	一條	1-1	0.5	1	0	0	.0028
NR	yesterday	Lh225	Na	Bi14	魚	1-1	0	0	0	0	.0023
VB	caught	De098	V+Di	Hm05	捕到	1-1	0	1	0	0	.0010
NN	fish	Af100	Nd	Tq23	昨天	1-1	0	3	0	0	.0004
NN	fish	Ah120	Nd	Tq23	昨天	1-1	0	3	0	0	.0004
NN	fish	Ea017	Nd	Tq23	昨天	1-1	0	3	0	0	.0004
NN	fish	Eb031	Nd	Tq23	昨天	1-1	0	3	0	0	.0004
NN	fish	Ab032	Nd	Tq23	昨天	1-1	0	3	0	0	.0004
After aligning “yesterday” and “昨天”											
NN	fish	Ab032	Na	Bi14	魚	1-1	0.75	1	15.3	0.0017	.0159
PP	I	Gh280	Nh	Na02	我	1-1	1	1	0	0	.0076
PP	I	Gh280	Nh	Na05	我	1-1	1	1	0	0	.0076
NN	fish	Af100	Na	Bi14	魚	1-1	0.75	1	0	0	.0034
NN	fish	Ah120	Na	Bi14	魚	1-1	0.75	1	0	0	.0034
NN	fish	Ea017	Na	Bi14	魚	1-1	0.75	1	0	0	.0034
NN	fish	Eb031	Na	Bi14	魚	1-1	0.75	1	0	0	.0034
AT	a	Nd098	Ne	Qa04	一條	1-1	0.5	1	0	0	.0028
VB	caught	De098	V+Di	Hm05	捕到	1-1	0	1	0	0	.0010
After aligning “fish” and “魚”											
PP	I	Gh280	Nh	Na02	我	1-1	1	0	0	0	.0179
PP	I	Gh280	Nh	Na05	我	1-1	1	0	0	0	.0179
AT	a	Nd098	Ne	Qa04	一條	1-1	0.5	0	0	0	.0067
VB	caught	De098	V+Di	Hm05	捕到	1-1	0	0	0	0	.0023
After aligning “I” and “我”											
AT	a	Nd098	Ne	Qa04	一條	1-1	0.5	0	0	0	.0067
VB	caught	De098	V+Di	Hm05	捕到	1-1	0	0	0	0	.0023
After aligning “a” and “一條”											
VB	caught	De098	V+Di	Hm05	捕到	1-1	0	0	0	0	.0023

Table 12 *Various factors for connection candidates.*

ST: English sentence		*TT*: Chinese sentence		
Word	Sense code	Word		Sense code
I	Gh280	我	(wuo, I)	Na05
caught	De098	捕到	(bu-dao, catch)	Hm05
a	Nd098	一條	(yi-tiao, one)	Qa04
fish	Ab032	魚	(yu, fish)	Bi14
yesterday	Lh225	昨天	(zuotian, yesterday)	Tq23

Table 13 *The final alignment of Example (5e, 5c).*

4. Experiment

4.1 Experimental Results

In this section, we present the results of algorithms for word alignment. Roughly 25,000 bilingual example sentences from LecDOCE were used as the training data. The training data were used primarily to acquire rules by the greedy learner and to determine empirically probability functions related to various factors. The algorithm's performance was then tested on the outside data. The outside test was on a set of 416 sentence pairs from a book on English sentence patterns. We chose this test set because it contained a comprehensive group of fifty-five sets of typical sentence patterns.

No. Words	No. Matched	No. Correct	Coverage	Precision
2862	2524	2272	88.2%	90.0%

Table 14 *Experimental result of SenseAlign.*

Table 14 indicates that acquired lexical information and existing lexical information in a bilingual dictionary can supplement each other to produce optimum alignment results. The generality of the approach is evident from the high coverage (88.2%) and precision rates (90.0%).

4.2 Typological Analysis of Alignment Errors

This section thoroughly analyzes the alignment results from the experiments and, in particular, the data related to cases where the algorithms failed. The analytical results demonstrate the strength and limitations of the methods and suggest possible improvements of the algorithm.

Metaphorical Usage

Metaphorical expressions are often language dependent, thus giving rise to a connection target which is different from the relevant dictionary translations. For instance, by the metaphorical expression (6e), one does not mean that someone really has green fingers, only that he is good at gardening. This metaphorical implication will not get across with a literal translation.

(6e) He has green fingers.

(6c) 他 精於 園藝。
ta jien-yu yuan-yi
he good at gardening

Collocation

Collocation is another reason for the deviation of the connection target from the dictionary translation, leading to failure of *SenseAlign*. However, unlike other deviations, bilingual collocations are not easily to tackle using class-based rules. For instance, in example sentence (7e, 7c), "*give order*" is a collocation, and the translation for "*give*" in such a collocation is usually " 下 ". However, the applicability is too low to warrant a mapping from "*give*" to " 下 ". In any case, deriving a give-to- 下 mapping would be an over-generalization.

(7e) The officer is the one who gives the orders.

(7c) 這個 軍官 就 是 下 命令 的 人。
zhe-ge jiunguan jiou shi xia mingling de ren
This officer actually is giving order DE person

Four-morpheme Mandarin Idioms and Free Translations

Paraphrased translation is a major source of alignment failure. Due to various considerations, including style and cultural differences, the translator does not always translate literally on a word-by-word basis. Adding and deleting words is commonplace, sometimes resulting in free translation. Such translations obviously create problems for word alignment. A significant amount of free translation arises due to the use of 4-morpheme Mandarin idioms for stylistic considerations. For instance, the clause "*hit close to home*" in (8e) translates into the idiom " 一針見血 ", and "*completely off base*"

in (9e) translates into the idiom " 大錯特錯 ." Apparently, these free translations are beyond the reach of the proposed method.

(8e)Everyone felt that the speaker's remarks hit close to home.

(8c)	大家	都	覺得	演講者	的	話	很	是	一針見血。
	djia	doul	juede	yanjiangzhe	de	hua	hen	shi	yzhenjianxve
	everyone	all	feel	speaker	DE	word	very	is	to the point

(9e) Your idea is completely off base.

(9c)	你的	想法	大錯特錯。
	nde	xiangfa	dacuotecuo
	your	idea	utterly wrong

4.3 Quantitative Error Analysis by Part-of-speech

Now we will look at alignment failure from a different angle: the part-of-speech. The error analysis by part-of-speech is shown in Table 15. Note that the majority of errors come from common nouns, light verbs, adverbs and prepositions. Obviously, function words are much more language-dependent and, therefore, more difficult to align correctly. Closer examination shows that connections related to function words are often one-to-many or even many-to-many, adding to the difficulty of connecting them correctly. These observations indicate the necessity of treating each part-of-speech differently - a context-sensitive lexical translation model for light verbs or perhaps a more elaborate model of fertility for function words.

Source Words			Alignment Errors		
POS	# of Words	Percentage	# of Errors	% in POS	% in All Errors
n	645	25.6	52	8.1	20.6
v	547	21.7	83	15.2	32.9
adj	141	5.6	12	8.5	4.8
adv	120	4.8	23	19.2	9.1
conj	68	2.7	6	8.8	2.4
det	336	13.3	14	4.2	5.6
pron	293	11.6	14	4.8	5.6
prep	225	8.9	34	15.1	13.5
others	149	5.9	14	9.4	5.6

Table 15 *Error analysis by POS.*

5. Discussion

In this section, we will justify the use of machine-readable lexical resources and a class-based approach. Although it is always difficult to compare different methods directly, we can contrast *SenseAlign* with other works related to word alignment in terms of resource requirements and statistical estimation reliability.

5.1 A Class-based Approach to Exploiting Machine-readable Lexical Resources

The crux of the NLP problems lies in knowledge acquisition, which is widely recognized as a bottleneck in development of NLP technology. To avoid this knowledge acquisition bottleneck, researchers have recently switched from manual and qualitative approaches to the MRD-based and corpus-based approaches. However, word-level knowledge acquired from dictionaries or corpora offers limited coverage. Take word sense disambiguation, a specific NLP task, for example. Lesk (1986) described a word-sense disambiguation technique based on the number of overlaps between words in a dictionary definition and words in the local context of the word to be disambiguated. Weak performance (50-70%) was reported. Yarowsky's (1992) WSD approach based on Roget's categories is a step in the right direction. The author reported a 92% precision rate for automatic disambiguation of the instances of 12 words in the Grolier Enclopedia.

The problem of word alignment is no exception. We believe that our proposed algorithm addresses the above problem by exploiting existing thesauri in addition to MRDs and corpora. The corpora provide us with training and testing materials, so that empirical knowledge can be derived and evaluated objectively. The thesauri provide a classification system that can be utilized to generalize the empirical knowledge gleaned from corpora. The approach of coupling corpora with thesauri to gain both empiricality and generality is broadly in line with the approaches used by Yarowsky (1992),Resnik and Hearst (1993), Utsuro, Uchimoto, Matsumoto and Nagao (1994), and Vanderwende (1994). The Vanderwende (1994) approach of using thesaurus-like information to interpret noun sequences is particularly of interest. Contrary to previous MRD-based works, an element of inference is added to word-for-word matching. The inference is realized through taxonomic relations, such as hyponyms and hypernyms extracted from LDOCE.

SenseAlign achieves a degree of generality since the elements of a word pair can be correctly aligned, even when they occur rarely or even once in the corpus. This kind of generality is unattainable by statistically trained word-based models. Class-based models obviously offer additional advantages of a smaller storage requirement and higher system

efficiency. Such advantages have their costs, for class-based models may be over-generalized and miss word-specific rules. However, class-based systems have produce results indicating that the advantages outweigh the disadvantages.

Obviously, *SenseAlign* is only one of many possible formulations of the class-based approach to word alignment using both a dictionary and thesaurus. Ker and Chang (1997) described a similar *ClassAlign* algorithm with a number of differences: 1. *ClassAlign* does not commit itself to a certain segmentation of Chinese sentences as *SenseAlign* does. 2. *ClassAlign* does not identify any morpho-synctical constructions in Chinese sentences as *SenseAlign* does. 3. Unlike *SenseAlign*'s repeated evaluation of distortion, *ClassAlign* calculates the distortion, once and for all, relative to anchors cast by the *DictAlign* algorithm. 4. *ClassAlign* selects alignment rules by balancing both applicability and specificity; thus, it does not increase coverage at the expense of precision. *ClassAlign* tends to make fewer commitment errors while *SenseAlign* tends to make fewer omission errors. Chen, Chang, Ker and Chen (1997) described a much simpler *TopAlign* algorithm which does not require segmentation of Chinese sentences. *TopAlign* takes advantage of various clusters based on a source language thesurus, LLOCE, instead of one thesurus for each of the two languages as in the case of *SenseAlign* and *ClassAlign*. Experimental results show that *TopAlign* runs much faster.

5.2 Other Methods Based on Mutual Information, X^2-like Statistics, and Frequency

Gale and Church (1991) showed through a near-miss example that ϕ^2, a X^2-like statistic, works better than mutual information for selecting strongly associated word pairs for use in word alignment. In their study, they contended that the X^2-like statistic works better because it uses co-nonoccurrence and the two off-diagonal values of the contingency table (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. Their results indicate that although precision is improved, coverage is not higher than that of other word-based approaches.

Focusing on improving coverage, we have chosen to use frequency coupled with simple filtering according to fan-out in the acquisition of class-based rules. Rules that provide the most instances of plausible connection are selected. Our approach differs from those based on a word-specific, mutual information-like statistic that select strongly associated word pairs which may have a weak presence in the data. The experimental results confirm the findings of several recent works on terminology extraction and structural disambiguation. Daille (1994) demonstrated that simple criteria related to

frequency coupled with a linguistic filter work better than mutual information for terminology extraction. Justeson and Katz (1995) also gave experimental results supporting a similar finding. Recent work involving structural disambiguation (Alshawi and Carter, 1994; Brill and Resnik, 1994) has also indicated that statistics related to frequency outperform mutual information and the X^2 statistic.

6. Concluding Remarks

This paper has presented an algorithm capable of identifying words and their translations in a bilingual corpus. It is effective for specific linguistic reasons. A significant majority of words in bilingual sentences have diverging translations; those translations are not often found in a bilingual dictionary. However, these deviations are largely limited within the classes defined in thesauri. Therefore, by using a class-based approach, the problem's complexity can be reduced. The results of experiments in this study have demonstrated that the method provides hefty coverage and precision rates well over 85%. In general, a small amount of precision can apparently be sacrificed to gain a substantial increase in coverage.

The algorithm's performance discussed here can definitely be improved by enhancing the various components of the algorithm, e.g., morphological analyses, a bilingual dictionary, monolingual thesauri, and rule acquisition. However, this work has presented a workable basis for processing bilingual corpus. This has wide implications for a variety of language tasks ranging from the obvious, machine translation and word sense disambiguation, to the unexpected, second language acquisition.

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 this is true, the algorithm presented here should be adaptable to other language pairs. The prospects for Japanese, in particular, seem highly promising. Work on alignment of English-Japanese texts using both dictionaries and statistics has been described by Matsumoto, Ishimoto and Utsuro (1993) and Utsuro, Ikeda, Yamane, Matsumoto and Nagao (1994).

There are a number of exciting future directions for continuing this work, including: (1) adding an automatic preprocessing step, sentence alignment, (2) representing the alignment results at the structural or symbolic level, (3) applying the result of alignment to statistical or hybrid machine translation systems.

Acknowledgments

The authors would like to thank the National Science Council of the Republic of China for financial support of this work under Contract No. NSC 82-0408-E-007-195. We would like to thank Liming Yu at Zebra English Service Union and Betty Teng and Nora Liu at Longman Asia Limited for making machine readable dictionaries available to us. Special thanks are due to Mathis H. C. Chen for preprocessing work on the MRD. Thanks are also due to Keh-Yih Su for many helpful comments. We are also thankful to the anonymous reviewers for many useful suggestions.

References

Alshawi, H. and D. Carter, "Training and Scaling Preference Functions for Disambiguation," *Computational Linguistics*, 20:4, 1994, pp. 635-648.

Brill, E. and P. Resnik, "A Rule Based Approach to Prepositional Phrase Attachment," In *Proceedings of the 15th International Conference on Computational Linguistics*, 1994, pp. 1198-1204, Kyoto Japan.

Brill, E., "A Simple Rule-Based Part of Speech Tagger," In *Proceedings of the third Conference on Applied Natural Language Processing*, 1992, pp. 152-155, ACL, Trento, Italy.

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

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, "A Statistical Approach to Machine Translation," *Computational Linguistics*, 16:2, 1990, pp. 79-85.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, 19:2, 1993, pp. 263-311.

Chao, Y. R., *A Grammar of Spoken Chinese*, University of California Press, 1968.

Chen, J. N. and J. S. Chang, "Towards Generality and Modularity in Statistical Word Sense Disambiguation," In *Proceeding of 2nd Pacific Asia Conference on Formal and Computational Linguistics*, 1994, pp. 45-48.

Chen, M. H. C. and J. S. Chang, "Structural Ambiguity and Conceptual Information Retrieval," In *Proceeding of 10th Pacific Asia Conference on Language, Information and Computation, 1995*, pp. 115-120, Hong Kong.

Chen, M. H. C., J. S. Chang, S. J. Ker, and J. N. Chen, "TopAlign: Word Alignment for Bilingual Corpora Based on Topical Clusters of Dictionary Entries and Translations," In, *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine*

Translation, 1997, pp. 127-134.

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

Church, K. W., "Char-Align: A Program for Aligning Parallel Text at the Character Level," In *Proceedings of the 31st Annual Meeting of Association for Computational Linguistics*, 1993, pp. 1-8.

Dagan, I., K. W. Church and W. A. Gale, "Robust Bilingual Word Alignment for Machine Aided Translation," In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, 1993, pp. 1-8.

Daille, B., "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology," In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, Workshop at the 32nd Annual Meeting of the ACL, 1994, pp. 29-36.

Dempster, A., N. Laird, and D. Rubin, "Maximum Likelihood from incomplete data via the EM Algorithm," *Journal of the Royal Statistical Society*, 39(B), pp. 1-38.

Dice, L. R. "Measures of the Amount of Ecologic Association between Species," *Journal of Ecology*, 1945, 26: 297-302.

Fujii, H. and W. B. Croft, "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*, 1993, pp. 237-246.

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

Fung, P. and K. McKeown, "Aligning Noisy Parallel Corpora Across Language Groups: Word Pair Feature Matching by Dynamic Time Warping," In *Technology Partnerships for Crossing the Language Barrier, Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 1994, pp. 81-88, Columbia, Maryland, USA.

Gale, W. A. and K. W. Church, and Yarowsky, "Using Bilingual Materials to Develop Word Sense Disambiguation Methods," In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, 1992, pp. 101-112, Montreal, Canada.

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

Justeson, J. S. and S. M. Katz, "Technical Terminology: Some Linguistic Properties and An

Algorithm for Identification in Text," *Natural Language Engineering*, 1:1, 1995, pp. 9-27, Cambridge University Press.

Kay, M. and M. Roscheisen, "Text-Translation Alignment," *Computational Linguistics*, 19:1, 1993, pp. 121-142.

Ker, S. J. and J. S. Chang, "A Class-base Approach to Word Alignment," *Computational Linguistics*, 1997, 23:2, pp. 313-343.

Ker, S. J. and J. S. Chang, "Aligning More Words with High Precision for Small Bilingual Corpora," In *Proceedings of the 16th International Conference on Computational Linguistics*, 1996, pp. 210-215, Copenhagen, Danmark.

Lesk, M. E., "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," In *Proceedings of the ACM SIGDOC Conference*, 1986, pp. 24-26, Toronto, Canada.

Longman Group, *Longman English-Chinese Dictionary of Contemporary English*, Published by Longman Group (Far East) Ltd., 1992, Hong Kong.

Matsumoto, Y., H. Ishimoto and T. Utsuro, "Structural Matching of Parallel Texts," In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 1993, pp. 1-30, Ohio, USA.

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

Mei, J. J., I. M. Zhu, Y. C. Gao and H. S. Yin, *Tongyici Cilin* (Word forest of synonyms), Tong Hua Publishing, 1993, Taipei, (traditional Chinese edition of a simplified Chinese edition published in 1984).

Proctor, P., *Longman English-Chinese Dictionary of Contemporary English*, Longman Group (Far East), 1988, Hong Kong.

Resnik, P., and M. A. Hearst, "Structure Ambiguity and Conceptual Relations," In *Proceedings of the Third Conference on Applied Natural Language Processing, Association for Computational Linguistics*, ACL, 1993, pp. 104-110.

Smadja, F., K. R. McKeown, and V. Hatzivassiloglou, "Translating Collocations for Bilingual Lexicons: A Statistical Approach," *Computational Linguistics*, 1996, 22:1, pp. 1-38.

Utsuro, T., H. Ikeda, M. Yamane, M. Matsumoto, and M. Nagao, "Bilingual Text Matching Using Bilingual Dictionary and Statistics," In *Proceedings of the 15th International Conference on Computational Linguistics*, 1994, pp. 1076-1082, Kyoto Japan.

Utsuro, T., K. Uchimoto, M. Matsumoto, and M. Nagao, "Thesaurus-Based Efficient Example Retrieval by Generating Retrieval Queries from Similarities," In *Proceedings of the 15th*

International Conference on Computational Linguistics, 1994, pp. 1044-1048, Kyoto Japan.

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

Wu, D. and X. Xia, "Learning an English-Chinese Lexicon from a Parallel Corpus," In *Proceeding of the first Conference of the American Machine Translation Association: Technology Partnerships for Crossing the Language Barrier*, 1994, pp. 206-213, Columbia, Maryland, USA.

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

Appendix A Description for CILIN

In CILIN, the categories are organized as a conceptual ontology of three levels: gross categories, intermediate categories and detailed categories. In Table A-1, each of the twelve gross categories is denoted by an upper case letter.

Letter	Gross Category	Letter	Gross Category
A	Human	B	Animate/Inanimate
C	Time/Space	D	Abstract
E	Characteristics	F	Body motion
G	Cognition verb	H	Social activities
I	State Verb	J	Relation verb
K	Function words	L	Greetings

Table A-1 *The gross categories of CILIN.*

A gross category has from 1 to 19 intermediate categories listed under it. In total, there are 94 intermediate categories. Under each intermediate category, there are from 5 to 55 detailed categories, and 2 digits are used to represent detailed categories. There are 1428 detailed categories. Two examples of detailed categories are listed as Table A-2.

Category	Meanings	Examples
Dm05	Schools	xueixiao (school), zhongxuei (high school), sifan (teacher's college)
Hf03	drive, navigate	xianshi (drive), hangxing (navigate), fachuang (row a boat)

Table A-2 Two detailed categories and their related words from CILIN.

From the above description, we can see that words are organized mostly according to their semantic properties. Thus, personal Pronouns such as 我 (wuo, I), are listed under "A" categories alone with content nouns like 人民 (ren-ming, people). This would cause problems for the task of word alignment. Therefore, we have identified places where function words and content words are listed under the same category and given those function words a different code:

(1) Personal Pronouns were given new categories *Na*, *Nb*, and *Nc* to distinguish them from the content words listed under *Aa*.
(2) Function words related to quantity, number, and measurement were taken out of the intermediate category *Dn* and given intermediate categories *Qa*, *Qb*, and *Ma*, respectively.
(3) The gross category *C* was split into two new gross categories *T* and *L* for time and location, respectively.
(4) To broaden the coverage of CILIN, we have also added words into detailed categories of CILIN using an automatic procedure which exploits the so-called *thesaury effect* of Chinese characters and Kanji (Fujii and Croft 1993). Some examples of added words are shown as Table A-3.

Chinese Word	Category	Chinese Word	Category
州長 (zhouzhang, governor)	Af10	鍵 (jian, key)	Bp13
次長 (cizhang, deputy)	Af10	雞毛撢子 (jimaodanzi, duster)	Bp13
車長 (cezhang, conductor)	Af10	鐵釘 (tiedien, nail)	Bp13
典獄長 (dianyuzhang, warden)	Af10	鐵棒 (tieban, iron rod)	Bp13

Table A-3 *Some added words and their categories from CILIN.*

Appendix B Description for LLOCE

Unlike CILIN, the categories of LLOCE are organized primarily according to subject matter. In the first level, fourteen major subjects are denoted with reference letters from A to N as shown in Table B-1.

Ref. Letter	Subject
A	Life and living things
B	Body; it function and welfare
C	People and family
D	Building, houses, home, clothes, belongings, and personal care
E	Food, drink, and Farming
F	Feeling, emotions, attitudes, and sensations
G	Thought, communication, language, and grammar
H	Substance, materials, objects, and equipment
I	Arts/Crafts, science/technology, industry/education
J	Number, measurement, money, and commerce
K	Entailment, sports and games
L	Space and time
M	Movement, location, travel, and transportation
N	General and abstract terms

Table B-1 *Subjects and their reference letters from LLOCE.*

There are from 7 to 12 titles listed under a subject. For instance, the titles listed under subject *H* are shown in Table B-2 to illustrate.

Ref. No.	Code	Title
H1-	a	Substances and materials in general
H30-	b	Objects generally
H60-	c	Specific substances and material
H110-	d	Equipment, machines, and instruments
H140-	e	Tools
H170-	f	Containers
H200-	g	Electricity and electrical equipment
H230-	h	Weapons

Table B-2 *The titles of subjects H and their reference letters from LLOCE.*

Under each title, there are from 10 to 50 sets of related words. Each set is given a 3-digit reference number. The titles are not reflected in the original LLOCE reference code. In order to represent this implicit grouping, we have assigned a lower case letter to each title. For example, "*objects generally*" is denoted using the letter *b*, and the reference code *H030* is replaced with *Hb030*. Therefore, each set is denoted by a upper case SUBJECT letter, a lower case TITLE letter and a 3-digit SET number. There are 2504 sets in total. Some sets from LLOCE are listed in Table B-3.

Ref. code	Meaning	Related words
Gb030	knowing and being conscious	recognize, be aware, be conscious of
Jf130	selling and butying	sell, retail, realize, market, buy, purchase, acquire, get, pawn, treat, patronize

Table B-3 *Some sets and their related words from LLOCE.*