A Class-based Approach to Word Alignment

Sue J. Ker*
National Tsing Hua University

Jason S. Chang*
National Tsing Hua University

This paper presents an algorithm capable of identifying the translation for each word in a bilingual corpus. Previously proposed methods rely heavily on word-based statistics. Under a word-based approach, frequent words with a consistent translation can be aligned at a high rate of precision. However, words that are less frequent or exhibit diverse translations generally do not have statistically significant evidence for confident alignment, thereby leading to incomplete or incorrect alignments. The algorithm proposed herein attempts to broaden coverage by exploiting lexicographic resources. To this end, we draw on the two classification systems of words in Longman Lexicon of Contemporary English (LLOCE) and Tongyici Cilin (Synonym Forest, CILIN). Automatically acquired class-based alignment rules are used to compensate for what is lacking in a bilingual dictionary such as the English-Chinese version of the Longman Dictionary of Contemporary English (LecDOCE). In addition, this alignment method is implemented using LecDOCE examples and their translations for training and testing, while further examples from a technical manual in both English and Chinese are used for an open test. Quantitative results of the closed and open tests are also summarized.

1. Introduction

Brown, Cocke, Della Pietra, Della Pietra, Jelinek, Lafferty, Mercer, and Roosin (1990) advocate a statistical approach to machine translation (MT) and initiate much of the recent interest in bilingual corpora. Statistical machine translation (SMT) can be understood as a word-by-word model consisting of two submodels: a **language model** for generating a source text segment S and a **translation model** for mapping S to its translation T. They recommend using a bilingual corpus to train the parameters of **translation probability**, $\Pr(S \mid T)$ in the translation model. For MT and other purposes, many methods have been proposed for sentence alignment of the Hansards, an English-French corpus of Canadian parliamentary debates (Brown, Lai, and Mercer 1991; Gale and Church 1991a; Simard, Foster, and Isabelle 1992; Chen 1993; Gale and Church 1993), and for other language pairs, including English-German, English-Chinese, and English-Japanese (Kay and Röscheisen 1993; Church, Dagan, Gale, Fung, Helfman, and Satish 1993; Fung and McKeown 1994; Wu 1994). Alignment at other levels of resolution is obviously useful. A section, paragraph, sentence, phrase, collocation, or word can be aligned to its translation (Kupiec 1993; Smadja, McKeown, and Hatzivassiloglou 1996). Other logical approaches involve aligning parse trees of a sentence and its translation (Matsumoto, Ishimoto, and Utsuro 1993; Meyers, Yangarber, and Grishman 1996), or simultaneously generating parse trees and alignment arrangements (Wu 1995).

* Department of Computer Science, National Tsing Hua University, Hsinchu, 30043, Taiwan, ROC. E-mail: ksj@volans.cs.scu.edu.tw; jschang@cs.nthu.edu.tw

In addition to machine translation, many applications for aligned corpora have been suggested, including machine-aided translation (Shemtov 1993), translation assessment and critiquing tools (Isabelle 1992; des Tombe and Armstrong-Warwick 1993; Macklovitch 1994), text generation (Smadja 1992; Smadja, McKeown, and Hatzivassiloglou 1996), bilingual lexicography (Klavans and Tzoukermann 1990; Church and Gale 1991; Daille, Gaussier, and Lange 1994; Kupiec 1993; van der Eijk 1993; Li 1994; Wu and Xia 1994), and word-sense disambiguation (Gale, Church, and Yarowsky 1992; Chang, Chen, Sheng, and Ker 1996). For these applications, we must go one step further from sentence alignment and identify alignment at the word level. In the process of word alignment, the translation of each source word is identified. This study concentrates primarily on identifying alignment at the word level for a given sentence and its translation.

In the context of SMT, Brown et al. (1993) present a series of five models of $\Pr(S \mid T)$ for word alignment. Model 1 assumes that $\Pr(S \mid T)$ depends only on **lexical translation probability** (LTP) $t(s \mid t)$, that is, the probability that the ith word s in S translates into the jth word t in T. The pair of words (s,t), or more precisely (s,t,i,j) since there could be more than one instance of s or t, is called a **connection**. Model 2 enhances Model 1 by considering the dependence of $\Pr(S \mid T)$ on the **distortion probability** (DP) $d(i \mid j,l,m)$ where l and m are the respective lengths of S and T measured in number of words. Brown et al. (1990) propose using an adaptive Expectation and Maximization (EM) algorithm to estimate the parameters for LTP and DP from a bilingual corpus. The EM algorithm iterates between two phases to estimate LTP and DP until both functions converge. In the **expectation phase**, the parameters $t(s \mid t)$ and $d(i \mid j,l,m)$ in the SMT model for all possible values of s, t, i, j, l, and m are estimated from the sample of an aligned bilingual corpus. In the **maximization phase**, each sentence-translation pair in the corpus is aligned by maximizing the translation probability, $\Pr(S \mid T)$. They examine the feasibility of aligning the English-French Hansards corpus using the SMT model, on both the sentence level and the word level. The SMT model is then tested for the task of machine translation. The model produces 35 acceptable translations for 73 sentences. However, to our knowledge, the degree of success of word alignment has not yet been explored.

Dagan, Church, and Gale (1993) observe that reliably distinguishing sentence boundaries for a noisy bilingual text scanned by an OCR device is quite difficult. In such a circumstance, they recommend aligning words directly without the preprocessing phase of sentence alignment. Under that proposal, a rough character-by-character alignment is first performed. Based on the character alignment, words are subsequently aligned based on a modified version of Brown et al.'s Model 2. The authors report that 60.5% of 65,000 words in a noisy document are correctly aligned. For 84% of the words, the offset from correct alignment is at most 3.

Gale and Church (1991b) present an alternative algorithm that does not estimate and store probabilities for all word pairs to reduce memory requirement and to ensure robustness of probability estimation. Instead, for each source word s, only a handful of target words strongly associated with s are found and stored. Such a task is achieved by applying a χ^2-like statistic. They report that the method produces highly precise (95%) alignment for 61% of the words in the 800 sentences tested.

This paper is motivated by the following observations: First, the above survey clearly reveals that word-based methods offer only limited coverage even after they are trained with an extremely large bilingual corpus. Second, we believe that for most applications, low coverage is just as serious 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.

This paper presents a word alignment algorithm based on classification in existing thesauri. The proposed algorithm, called *ClassAlign*, relies on an automatic procedure to acquire class-based alignment rules; it does not employ word-by-word translation probabilities, nor does it use an iterative EM algorithm for estimating such probabilities. Experimental results indicate that classification based on existing thesauri is highly effective in broadening coverage while maintaining a high precision rate.

The rest of this paper is organized as follows: In Section 2 we briefly discuss the nature of text and translation that justifies a class-based approach. A set of three algorithms leading to class-based alignment are outlined in Section 3. The algorithms' effectiveness is demonstrated through examples and their translations in the LecDOCE (Longman Group 1992), a bilingual version of the Longman Dictionary of Contemporary English (LDOCE, Proctor 1988), as well as sentences from bilingual texts in the LightShip User's Guide (Pilot Software Inc. 1993; Galaxy Software Services 1994). The experiments we undertook to assess the performance of these algorithms are the topic of Section 4. Quantitative experimental results are also summarized. In Section 5, we analyze the experimental results and consider ways in which the proposed algorithms might be extended and improved. Concluding remarks are made in Section 6.

2. Text and Translation as a Class-to-Class Mapping

The discussion in Section 1 indicates the limitations of statistical methods. As an alternative, we examine the feasibility of using an everyday bilingual dictionary in machine-readable form for word alignment. With tens of thousands of headword-and-translation pairs that can be used to propose high-precision connections, a bilingual machine-readable dictionary (MRD) surprisingly leads to even lower coverage than a statistically-derived lexicon. Below, observations are made to account for the reason why a substantial portion of translations deviate from what is listed in the bilingual MRD or what is statistically probable. Such deviations inhibit word-based methods from achieving broad coverage. We contend that a word's translational deviation is mostly bounded within the relevant semantic classes, thus justifying a class-based approach to word alignment.

2.1 Diverse In-Context Translations

Given that the translations for a headword (dictionary translations, DTs for short) can be extracted from a bilingual MRD such as the LecDOCE, a word in *S* can be aligned at a high precision rate with its DTs found in *T*. Headword-and-translation pairs are a reliable knowledge source for word alignment. However, they cover only a small part of the connections in an average sentence and its translation. Our experiments reveal that the translations of a word in context (in-context translations, ICTs for short) are frequently more diversified than the offerings in an everyday bilingual dictionary. More specifically, less than 30% of the English words in the context of LecDOCE examples translate into one of the relevant DTs in the same dictionary.

Translations in an everyday dictionary are meant to provide the reader with the idea of what is implied by the headword *out of context*; they are frequently more of an *explanation* than a *translation*. For instance, one LecDOCE sense entry defines the word *boy* as 'infml esp. AmE a male person, of any age, from a given place' and gives 某地來之男人 (*modi lai zhi nanren*) as the translation relevant to this particular sense. Such a 'translation' per se seems unlikely to appear as the ICT of *boy*. Aside from this fundamental difference, behind the disparity between DT and ICT are a plethora of factors. These include (1) a failure on the dictionary's (or the statistically derived lexicon's) part to cover a needed word sense, (2) mismatches in sense specificity between

the two languages, (3) collocation pattern, and (4) frequent use of interchangeable synonymous translation not covered in the dictionary. These factors are the reasons why many translations are statistically unlikely, thereby leading to a low coverage rate for word-based methods.

Sense Gaps. Dictionary and statistically derived translations might not cover the word sense appropriate to a given word in context. This is particularly true when using an everyday dictionary for aligning bilingual technical manuals. For instance, the LecDOCE lists four senses and relevant translations for the word *click*, including (1) 發出輕微而短促的聲音 'to make a slight and short sound', (2) 成功 'to succeed', (3) 使恍然大悟 'to fall into place', (4) 情投意合 'to be a quick success, esp. with members of the opposite sex', none of which is the right sense for 按 'to press', the translation of *click* in the context of (E1):

(E1) Click anywhere else on the screen background or press ESC.

(C1) 在畫面背景的其他地方按一下或按 ESC。

Mismatch in Sense Specificity. Dictionary treatment of word senses in the source language might not correspond to the level of specificity for the relevant concepts in the target language. For instance, the LecDOCE differentiates two word senses for the word *news* by the means in which it is reported: whether it is via electronic (radio or television) or non-electronic (newspaper) media. In Chinese, the relevant concept is also differentiated according to how it is reported; however, the difference is between mass media (translated as 新聞) and personal communication (translated as 消息). The following examples (E2, C2) and (E3, C3) demonstrate this particular instance of mismatch in sense specificity.

(E2) to listen to the 7 o'clock news broadcast.

(C2) 收聽七點鐘的新聞廣播。

(E3) Our latest *news* of our son was a letter a month ago.

(C3) 有關兒子的最新消息是從我們一個月前收到的信得到的。

Collocation Pattern. The collocation pattern often forces the choice of an ICT quite different from the DTs. For instance, the LecDOCE lists 新聞 'news' and 新聞報導 'news reportage' as the translations for *news*. However, the translation for *news* modified by *bad* is usually 消息 'message'. Similarly, *lady* modified by *old* almost surely translates into 太太 'wife' rather than the DTs, 女士 'lady' or 女性 'woman' given in the LecDOCE. The following examples provide further details.

(E4) Nothing but bad news in the newspaper today.

(C4) 今天報紙上只有壞消息。

(C4′) *今天報紙上只有壞新聞。

(E5) He was very attentive to the old lady and did everything for her.

(C5) 他對那位老太太非常關心，什麼事都幫她做。

(C5′) *他對那位老女士非常關心，什麼事都幫她做。

(C5″) *他對那位老女性非常關心，什麼事都幫她做。

(E6) He abdicated all responsibility for the care of the child.

(C6) 他放棄了照顧該小孩的一切責任。

(C6′) *他放棄了照料該小孩的一切責任。

Interchangeable Synonymous Translation. Disparity might arise simply because an interchangeable synonym of a DT is used. For instance, for the following LecDOCE examples, the synonyms 捕到, 遇見, and 悲痛 are present in the translations, instead of the respective DTs, 打到, 遇到, and 悲傷. If these ICTs in (C7, C8, C9) are replaced by the DTs, the subsequent translations (C7′, C8′, C9′) remain correct.

(E7) I caught a fish yesterday.

(C7) 昨天我捕到一條魚。

(C7′) 昨天我打到一條魚。

(E8) I have never met so nice a girl.

(C8) 我從未遇見過這麼好的女孩子。

(C8′) 我從未遇到過這麼好的女孩子。

(E9) He abandoned himself to grief.

(C9) 他不勝悲痛。

(C9′) 他不勝悲傷。

A statistically-derived lexicon generally fares better than MRDs in covering such synonymous translations. However, this is limited to synonyms that appear as an alternative translation frequently and consistently in a bilingual corpus.

Bounding the ICTs. An ICT may deviate from the relevant DTs for a variety of reasons, but the deviation is not without constraints. Table 1 lists some examples of deviating translations taken from the LecDOCE. Examples include the words *news, meet, lady, grief, care,* and *child* and their respective translations 消息, 遇見, 太太, 悲痛, 照顧, and 小孩. Notice that most in-context and dictionary translations of source words are bounded within the same category in a typical thesaurus such as the LLOCE (McArthur 1992) and CILIN (Mei et al. 1993). For instance, *Da19*-class words in CILIN (news and messages), 消息, 新聞, 報告 appear as the translations (DTs and ICTs) for *Ge194*-words in LLOCE (information and news) such as *news* and *report*. Similarly, *Id18*-class words (hitting, touching, meeting, and missing), 遇見, 遇到, 偶遇, 邂逅 appear as the translations for *Mc072*-words (meeting people and things) such as *meet* and *encounter*. This finding suggests that LTP can be estimated more robustly via class-to-class mapping. Furthermore, such ICTs and DTs are often synonymous compounds

Table 1
Disparity between ICT and DT is bounded within thesaurus categories.

Example Sentences	Word	ICT	DT
What wonderful news: the painting on my wall is a Rembrandt!	news	消息	新聞
真是大好消息我牆上那幅畫是倫勃朗真蹟。	(Ge194)	(Da19)	(Da19)
Reports that the general is to be dismissed are gaining currency among government ministers.	report	消息	報告
將軍即將被解職的消息在政府部長間逐漸流傳開來。	(Ge194)	(Da19)	(Da19)
I have never met so nice a girl.	meet	遇見	遇到
我從未遇見過這麼好的女孩子。	(Mc072)	(Id18)	(Id18)
He encountered many difficulties.	encounter	遇到	偶遇; 邂逅
他遇到很多困難。	(Mc072)	(Id18)	(Id18)
He was very attentive to the old lady and did everything for her.	lady	太太	女士
他對那位老太太非常關心, 什麼事都幫她做。	(Ca005)	(Ab01)	(Ab01)
She's a very wealthy woman, and moves in the highest circles of society.	woman	女士	女性; 婦女; 女流; 女人
她是位很有錢的女士, 活躍於高級社交圈。	(Ca002)	(Ab01)	(Ab01)
He abandoned himself to grief.	grief	悲痛	悲傷
他不勝悲痛。	(Fd082)	(Ga01)	(Ga01)
The sad man was in an abyss of hopelessness.	sad	悲傷	傷心; 悲慘; 哀愁; 難過
那悲傷的人正處在失望的深淵裏。	(Fd080)	(Ga01)	(Ga01)
He abdicated all responsibility for the care of the child.	care	照顧	照料
他放棄了照顧該小孩的一切責任。	(Nl366)	(Hi37)	(Hi37)
We should advertise for someone to look after the garden.	look after	照料	照顧
我們應登廣告招聘人來照料花園。	(Nf162)	(Hi37)	(Hi37)
He abdicated all responsibility for the care of the child.	child	小孩	嬰兒
他放棄了照顧該小孩的一切責任。	(Ca003)	(Ab04)	(Ab04)
John Smith? Yes - he's a local boy, I believe.	boy	人	男人
約翰史密斯? 是的, 我想他是本地人。	(Ca002)	(Ab02)	(Ab01)

that share a common morpheme. For instance, the (ICT, DT) pairs, (悲痛 and 悲傷) and (女士 and 女性) share a common morpheme 悲 'sad' and 女 'female', respectively. Fujii and Croft (1993) also point out a similar **thesaurus effect** of Mandarin morphemes in Japanese information retrieval (IR).[1]

Dictionary-based Alignment. The above observations suggest that a DT-based algorithm, coupled with morpheme-level partial matching, can be adopted to obtain a substantial

1 Fujii and Croft observe that a document is likely to be relevant if it contains an index term that has a morpheme (kanji) in common with a query term. More often than not, the index term and the query term are synonyms that might appear under the same category in a thesaurus. The authors call this phenomenon the thesaurus effect of kanji.

Table 2
Complete and partial matches against dictionary translations.

Example Sentences and Translations	Complete Matches (Headword, DT=ICT)	Partial Matches (Headword, DT, ICT)
I only know it was a dog and not a cat that bit me. 我只知道咬我的是狗不是貓。	(know, 知道), (bit, 咬), (dog, 种), (cat, 猫)	(only, 只有, 只)
I have made you an absolute promise that I will help you. 我已經無條件地答應要幫助你。	(help, 幫助), (you, 你), (will, 要)	(have, 已, 已經), (absolute, 無條件的, 無條件地)
There was an acute lack of food. 嚴重缺乏食物。	(lack, 缺乏), (food, 食物)	(acute, 嚴重的, 嚴重)
He added the wood to the fire. 他給火添了木柴。	(he, 他), (fire, 火)	(wood, 木材, 木柴) (wood, 柴火, 木柴)

number of high-precision connections. Experimental results indicate that a DT-based method connects over 40% of words in LecDOCE examples with their ICTs using this rudimentary method. Table 2 presents some examples from the experiments, indicating the connections that are attributed to a complete or partial match using headword-and-DT pairs extracted from the LecDOCE. For instance, partial match enables the method to pair up *only* and 只 according to its DT 只有. These connections can be subsequently used as the basis for generalizing to a class-based alignment rule in the form of (X, Y), which stipulate the connection between an X-class word and a Y-class word.

2.2 Class-based Word Alignment

To ensure broad coverage, the class-based approach seems to be a promising alternative to word-based methods. Classes can be formed from words in more than one way. Automatic statistical methods for derived classes (Brown, Della Pietra, deSouza, Lai, and Mercer 1992) are not appropriate, since they also suffer low coverage due to data sparseness. Classes formed from morphologically related words are easy to derive and apply. Morphological classes can be formed, either from words that start with the same five-character prefix as in Gale and Church (1991b), or rigorous analysis as suggested in Brown, Della Pietra, Della Pietra, Lafferty, and Mercer (1992). Although easily applicable, morphological classes are not particularly effective in broadening coverage of word alignment. Chang and Chen (1994) also examine the feasibility of using part-of-speech classes. A potential alternative involves adopting categories available in machine-readable lexicographic resources such as Roget's thesaurus (Chapman 1977) or hand-crafted computer lexicons (Miller, Beckwith, Fellbaum, Gross, and Miller 1990; McRoy 1992).

3. Algorithms Leading to Class-based Word Alignment

This section describes a series of three algorithms leading to a class-based system for word alignment. The first algorithm attempts to obtain reliable connections. The second algorithm generalizes the connections into a list of class-based rules, which stipulate that a pair of classes of words in the source and target languages are likely mutual translations. The third algorithm performs the actual word alignment based on the acquired rules, in addition to DTs.

3.1 Dictionary-based Word Alignment

This section describes a rudimentary word-alignment algorithm, *DictAlign*, based on the DTs from a bilingual MRD such as the LecDOCE. Consider a text and translation pair (S,T), a word s in S, and its ICT, t in T. Let DT_s denote the set of translations listed in the LecDOCE for the headword s. Recall that if for a word t in T, there is a dt in DT_s such that t matches dt completely or partially, then, t is likely to be the ICT of s. Taking advantage of this phenomenon, *DictAlign* computes the set $W_T = \{t \mid t \text{ is a word in } T\}$ and calculates the similarity between each t and the DT_s relevant to S. A similarity measure based on the unweighted Dice coefficient (Dice 1945) can be given as follows:

$$\text{Sim}(d,t) = \frac{2 \times |d \cap t|}{|d| + |t|} \tag{1}$$

where d,t = Mandarin morpheme strings,
$|d|$ = the number of the morphemes in d,
$|t|$ = the number of the morphemes in t,
$|d \cap t|$ = the number of the morphemes in the intersection of d and t.

Based on this similarity measure, the likelihood of a connection can be associated with the following formulation that links the likelihood of a connection to similarity with a DT:

$$\text{DTSim}(s,t) = \max_{d \in DT_s} \text{Sim}(d,t) \tag{2}$$

For instance, consider the following sentence and its Mandarin translation, focusing on the word *encounter*:

S = *He encountered many difficulties.*
T = 他遇到很多困難。

We will have the following:

W_S = {he, encounter, many, difficulty}
W_T = {他, 遇, 遇到, 很, 很多, 多, 困, 困難, 難}
$DT_{\text{encounter}}$ = {偶遇, 邂逅, 遭遇}.

Therefore, the connections relevant to *encounter* with nonzero DTSim values based on unweighted Dice coefficient are as follows:

$$\begin{aligned}
\text{DTSim}(\textit{encounter}, 遇) &= \max\{\text{Sim}(偶遇, 遇), \text{Sim}(邂逅, 遇), \text{Sim}(遭遇, 遇)\} \\
&= \max\left\{\frac{2\times1}{1+2}, \frac{0}{1+2}, \frac{2\times1}{1+2}\right\} = 0.67 \\
\text{DTSim}(\textit{encounter}, 遇到) &= \max\{\text{Sim}(偶遇, 遇到), \text{Sim}(邂逅, 遇到), \text{Sim}(遭遇, 遇到)\} \\
&= \max\left\{\frac{2\times1}{2+2}, \frac{0}{2+2}, \frac{2\times1}{2+2}\right\} = 0.5
\end{aligned}$$

The head morpheme in a word is usually more relevant in determining a word's meaning, just as content words carry more meaning than function words. Matching such a morpheme often implies a higher likelihood of finding the ICT. For instance, 遇 is the head morpheme of the DT 遇到, and should be given a heavier weight. Our

experiments indicate that by weighting morphemes, ICT ambiguity can be resolved more successfully. Assuming that such weights can be obtained in a manner similar to what is done in IR when assigning weights to index terms, the weighted Dice coefficient can be used by substituting weights for counts in equation (1) to arrive at the following:

$$\mathrm{Sim}(d,t) = \frac{2 \times |d \cap t|}{|d| + |t|} \tag{3}$$

where d, t = the Mandarin morpheme strings,
$|d|$ = total weights for the morphemes in d,
$|t|$ = total weights for the morphemes in t,
$|d \cap t|$ = total weights for the morphemes in the intersection of d and t.

The above descriptions are summarized as the *DictAlign* Algorithm:

Algorithm 1 (*DictAlign*) Align each word s in S with the ICT t in T based on DT_S.
Step 1: Remove all stop words in S to obtain a list of keywords, W_S.
Step 2: Lookup all possible words W_T of T in a dictionary.
Step 3: For each s in W_S, look up the root of s in a bilingual dictionary to obtain DT_S.
Step 4: For all $d \in DT_S$ and all $t \in W_T$, calculate $\mathrm{Sim}(d,t)$ according to equation (3).
Step 5: For each $(s,t) \in W_S \times W_T$, calculate $\mathrm{DTSim}(s,t)$ according to equation (2).
Step 6: For each word s, produce a connection (s,t), if $\mathrm{DTSim}(s,t)$ is maximized over $t \in W_T$ and $\mathrm{DTSim}(s,t) > h_1$ where h_1 is a preset threshold.
Step 7: Compile the list of connections and denote the list as *CONN*.

To illustrate how *DictAlign* works, consider the sentence pair (E10, C10). After the stopwords are removed, we obtain $W_S = \{old, lady, clad, fur, coat\}$. The list of words in T is also obtained by consulting a Chinese dictionary.[2] Subsequently, for each s in W_S, we lookup s in the LecDOCE to obtain DT_s. Table 3 shows dictionary translations relevant to (E10). Table 4 lists all W_S words along with the relevant DT_s, possible translation t, as well as the values of $\mathrm{Sim}(d,t)$ and $\mathrm{DTSim}(s,t)$. Table 5 displays the result *CONN* for various values of threshold h_1.

(E10) The old lady was clad in a fur coat.

(C10) 這位老婦人穿著皮裘。

3.2 Acquisition of Mutually Translatable Class Pairs

ClassAlign is conceived to capture the diversity of translations for broad-coverage alignment. One way to do so is via the classification of words in thesauri. More specifically, one can generalize from a connection (s,t) to a class-to-class mapping (X,Y) where X and Y are thesaurus classes containing s and t respectively. However, this simple intuition is complicated by the fact that a word might belong to more than one class, that is, if the classification is based on a thesaurus that allows for word-sense ambiguity. For a word in a particular context, if one considers classes that are not intended for the context, noise can be introduced. For instance, consider the

2 The dictionary used in this study is a combination of CILIN and an on-line dictionary developed by the CKIP group, Academy Sinica, Nankang, Taiwan.

Table 3
The DT_s for each word s in W_S for example (E10, C10).

Word	Root	Stopword	Dictionary Translations of Words in S
the	the	yes	None (all sense entries have an explanation in brackets; not a translation)
old	old	no	上年紀的, 以前的, 用舊的, 年老的, 年歲的, 老的, 認識許久的, ...
lady	lady	no	女士, 女子, 夫人, 女性, 貴婦之銜稱, 戀人, ...
was	be	yes	是, 存在, 在 ... 之上, 在 ... 周圍, ...
clad	clad	no	穿著, 覆蓋
in	in	yes	在 ... 之內, 在 ... 中, 在 ... 裏面, 在, 進入, ...
a	a	yes	一個, 一種, 一些, 一罐, 一瓶, 相同的, 同一的, ...
fur	fur	no	毛皮, 皮衣, 舌苔, 瓶垢, 獸皮之軟毛, 銹皮
coat	coat	no	大衣, 外衣, 外套, 表層, 動物之皮或毛

Table 4
All connection candidates and DTSim values in (E10, C10).

$s \in W_S$	$t \in W_T$	$dt \in DT_s$	Sim(dt,t)	DTSim(s,t)
old	老	年老的	0.54	0.74
old	老	老的	0.74	0.74
old	老婦	年老的	0.41	0.51
old	老婦	老的	0.51	0.51
old	老婦人	年老的	0.35	0.42
old	老婦人	老的	0.42	0.42
lady	老婦	貴婦之銜稱	0.30	0.30
lady	老婦人	貴婦之銜稱	0.27	0.31
lady	老婦人	戀人	0.30	0.31
lady	老婦人	夫人	0.31	0.31
lady	婦人	貴婦之銜稱	0.31	0.39
lady	婦人	戀人	0.37	0.39
lady	婦人	婦人	0.39	0.39
lady	婦	貴婦之銜稱	0.34	0.34
lady	人	戀人	0.52	0.56
lady	人	夫人	0.56	0.56
clad	穿	穿著	0.71	0.71
clad	穿著	穿著	1.00	1.00
clad	著	穿 著	0.62	0.62
fur	皮	獸皮之軟毛	0.31	0.67
fur	皮	銹皮	0.53	0.67
fur	皮	皮衣	0.67	0.67
fur	皮	毛皮	0.67	0.67
coat	皮	動物之皮或毛	0.31	0.31

Table 5
The results of running *DictAlign* on (E10, C10) under various thresholds.

Threshold = 0.7

Word s	Translation t	DTSim(s,t)
old	老	0.74
clad	穿著	1.00

Threshold = 0.67

Word s	Translation t	DTSim(s,t)
old	老	0.74
clad	穿著	1.00
fur	皮	0.67

Threshold = 0.5

Word s	Translation t	DTSim(s,t)
old	老	0.74
lady	人	0.56
clad	穿著	1.00
fur	皮	0.67

connection (*have*, 吃) found in the context of Example (E11, C11). According to the LLOCE, *have* belongs to the following topical sets: *Cb024* (relating to sex), *De081* (having and owning), *Ea003* (eating and drinking), *Li273* (auxiliary related to time), and *Nf159* (making necessary). The LLOCE class *Ea003* and CILIN class *Fc06* (to eat, chew, suck, and drink) are intended for this context. However, without that information, the following noisy rules might be introduced: (*Cb024*, *Fc06*), (*De081*, *Fc06*), (*Li273*, *Fc06*), (*Nf159*, *Fc06*), along with the signal (*Ea003*, *Fc06*).

(E11) Let's have breakfast early for a change.

(C11) 我們來個變化，早一點吃早餐吧。

The noise is usually distributed randomly while the signal tends to repeat itself. Nevertheless, connections (s,t) related to some ambiguous words s or t may cause noise to accumulate, leading to erroneous generalization. Therefore, one should try to throw away such noise. Moreover, any signal that gets thrown away by not considering (s,t) is often filled by a connection (s',t) where s' is a synonym of s. For instance, *get* is many ways ambiguous, as indicated in diversified ICTs in the LecDOCE examples in Table 6. However, each of these ICTs seems to form a connection with a less-ambiguous synonym of *get* such as *receive*, *reach*, and *understand* in LecDOCE examples. Table 6 provides further details.

As is typical in IR research, highly frequent and ambiguous words (known as stopwords in the IR literature) can be thrown out to reduce such noise. A list of stopwords used in the experiments includes the following:

a, at, be, drive, eye, field, fix, for, from, function, get, go, have, head, idea, in, into, lot, of, on, place, the, to, up, with, . . .[3]

3 Stopwords include *eye*, *field*, and *fix*, which are not usually member of stopword lists in IR. They are

Table 6
Some sentence-translation pairs containing the word *get*.

Example and Translation	Source/Class	ICT
I got a letter today. 我今天收到一封信。	get (*De083*, Getting and earning)	收到
It is unpleasant to receive anonymous letters. 收到匿名信令人不快。	receive (*De083*, Getting and earning)	收到
We got there at 8 o'clock. 我們八點鐘到達那裏。	get (*Ma005*, Arriving and reaching)	到達
His hunger was not appeased until he reached the hotel. 到達旅館之後他的飢腸才得到滿足。	reach (*Ma005*, Arriving and reaching)	到達
I don't get you; what do you mean? 我不明白，你是什麼意思？	get (*Gb031*, Understand and realize)	明白
I understood that it was time to leave. 我也就明白我該走了。	understand (*Gb031*, Understand and realize)	明白

With the difficulties of finding appropriate classification systems and suppressing noise now resolved, the question remains: How can class-to-class mapping be acquired? Just as with the derivation of a bilingual lexicon from a corpus, acquisition of such mapping requires a statistical measure. The Dice coefficient (Dice 1945) is a similarity measure that gauges the ratio of the members in one collection being identical to those of another collection. Smadja, McKeown, and Hatzivassiloglou (1996) propose to link co-occurrence to the Dice coefficient in their study of bilingual collocations. They observe that, unlike statistical measures related to mutual information, the Dice coefficient is insensitive to sample size and, thus, more effective for acquiring bilingual collocations from a bilingual corpus. Our experimental results confirm their observation. Under a formulation linking translation to conceptual similarity, the Dice coefficient is a very useful estimator of the class-to-class mapping.

Therefore, in this work, we measure the likelihood of class-to-class translation mapping in terms of the ratio of member pairs that are connections observed in a bilingual corpus. This ratio can be easily measured using the Dice coefficient as follows:

$$\text{ClassSim}(X,Y) = \frac{\sum_{a\in X} \text{From}(a,Y) + \sum_{b\in Y} \text{To}(X,b)}{|X| + |Y|} \tag{4}$$

where $|X|$ = the total number of the words in X,
$|Y|$ = the total number of the words in Y,
$\text{From}(a,Y)$ = 1, if $(\exists y \in Y)(a,y) \in ALLCONN$,
= 0, otherwise,

treated as such because of their high frequency in the LDOCE and their involvement in diverse LLOCE topics and CILIN categories. For instance, the LLOCE lists *eye* under the following topical sets: *Ac051* (part of an animal), *Aj151* (part of a plant), *Bc023* (part of human body), *Dg152* (part of a shoe), *Hd126* (part of a needle), etc.

$$\begin{aligned} \mathrm{To}(X,b) &= 1, \text{ if } (\exists x \in X)(x,b) \in ALLCONN, \\ &= 0, \text{ otherwise,} \\ ALLCONN &= \text{the word-translation pairs compiled from the results of running} \\ &\quad \textit{DictAlign} \text{ on all sentence-translation pairs of a bilingual corpus.} \end{aligned}$$

This naive estimator works efficiently for classes of compatible sizes. Occasionally, for extremely small or large classes, the coefficient does not accurately reflect how likely words in one class are to translate to words in another class. To remedy this problem, we explore the feasibility of weighting the member words. According to our result, weighting eradicates most instances of the problem caused by uneven classification. The weight assigned to each word should positively correlate to the frequency of the word so it reflects the expected ratio of word-translation pairs. Assuming that such weights can be obtained on the basis of each word's frequency in a bilingual corpus, weights can substitute for counts in equation (4) to arrive at the weighted version of the Dice coefficient shown below:

$$\mathrm{ClassSim}(X,Y) = \frac{\sum_{a \in X} \mathrm{From}(a,Y) + \sum_{b \in Y} \mathrm{To}(X,b)}{|X| + |Y|} \tag{5}$$

$$\begin{aligned} \text{where} \quad |X| &= \text{the total weights of the words in } X, \\ |Y| &= \text{the total weights of the words in } Y, \\ \mathrm{From}\ (a,Y) &= \text{the weight of } a, \text{ if } (\exists y \in Y)(a,y) \in ALLCONN, \\ &= 0, \text{ otherwise,} \\ \mathrm{To}(X,b) &= \text{the weight of } b, \text{ if } (\exists x \in X)(x,b) \in ALLCONN, \\ &= 0, \text{ otherwise,} \\ ALLCONN &= \text{the source-translation pairs obtained for all sentences and} \\ &\quad \text{translations in the training corpus using Algorithm 1.} \end{aligned}$$

Algorithm 2 summarizes the *ClassRule* algorithm for acquiring class-based rules. Table 7 presents a random sample of class-based rules acquired from connections found in LecDOCE examples and translations.

Algorithm 2. (*ClassRule*) **Acquisition of pairs of mutually translatable classes** (X, Y).

Step 1: Run *DictAlign* on the sentences in a bilingual corpus to obtain a list of initial connections *ALLCONN*.

Step 2: For all $X \in CX$ and all $Y \in CY$, compute similarity $\mathrm{ClassSim}(X,Y)$ based on weighted Dice coefficient given in (5), where CX and CY are some classification of words in the source and target languages, respectively.

Step 3: Produce an alignment rule (X, Y),

if $\mathrm{ClassSim}(X,Y) > h_1$, a preset threshold,
or if $\mathrm{ClassSim}(X,Y)$ is maximized over all X in CX or all Y in CY.

Step 4: Compile the list of such class pairs satisfying the conditions in Step 3 and denote the list as *RULES*.

Table 7
A random sample of mutually translatable classes.

ClassSim	LLOCE Class	CILIN Class	Gloss for LLOCE Classes	Gloss for CILIN Classes	Member Connections
0.92	Db025	Bn05	Ground, floors, and foundations	Foundations, ground, and ceilings	(floor, 地板) (ground, 地面)
0.85	Lh234	To21	Months of the year	Month	(April, 四月) (July, 七月)
0.75	Hf172	Bp07	Cups, plates, and bowls	Tray, cup, bowl, etc.	(bowl, 碗) (glass, 酒杯)
0.73	Ma031	Ih08	Moving faster and slower	Accelerate or decelerate	(decelerate, 減速) (speed up, 加速)
0.70	Cl231	An03	Kinds of thief	Robber, thief, etc.	(robber, 強盜) (thief, 賊)
0.70	Lc045	Bf05	Clouds, fog, and steam, etc.	Smoke, fog, etc.	(cloud, 煙霧) (fog, 霧)
0.62	Me122	Ln14	Edges, boundaries, and borders	Frontier, boundary, etc.	(border, 國境) (limit, 界限)
0.34	Eb030	Bh06	Vegetables in general	Vegetables	(bean, 豆) (onion, 洋蔥)
0.24	Ah124	Bk08	The feet of different creatures	Four limbs, arm, hand, and leg	(claw, 爪) (foot, 足)
0.07	Gf230	Dk11	Language, speech, dialect	Speech	(speech, 言詞) (tongue, 話)

Class (*Ca005*) = {Mrs, Ms, broad, dame, female, girl, lady, Madam, missis, miss, missus}

Class (*Ab01*) = {丈夫, 丫頭, 千金, 士女, 女人, 女人家, 女士, 女子, 女兒, 女性, 女郎, 子, 小姐,夫人, 太太, 少女, 少奶奶, 少婦, 奶奶, 母女, 先生, 男人, 男女, 男子, 男子漢, 兒女, 姑娘, 哥, 孫兒女, 婦人, 婦女, 貴婦人, 媳婦兒, ...}

From $(x, Ab01) = 1$, for all $x \in$ {dame, female, girl, lady, Madam, miss}

To $(Ca005, x) = 1$, for all $x \in$ {女人, 女, 女性, 女子, 女士, 小姐, 夫人}

$$\text{ClassSim}(Ca005, Ab01) = \frac{7+6}{11+92} = 0.13$$

Figure 1
The word classes *Ca005* and *Ab01* and their conceptual similarity.

The main step in *ClassRule* is illustrated through the calculation of the *ClassSim* value between LLOCE topical set *Ca005* (kinds of woman) and CILIN category *Ab01* (man and woman, 男人, 女人, 男女). For simplicity, the unweighted value of *ClassSim*(*Ca005*, *Ab01*) is calculated. *Ca005* and *Ab01* contain 11 and 92 words, respectively. In *ALLCONN*, six words in *Ca005*, i.e. *dame*, *female*, *girl*, *lady*, *Madam*, and *miss*, translate into words listed under *Ab01*. In the other direction, seven words listed under *Ab01*, i.e. 女人, 女, 女性, 女子, 女士, 小姐, and 夫人, are the translations from words listed under *Ca005*. Thus, the *ClassSim* value between *Ca005* and *Ab01* can be valued at $(7+6)/(11+92) = 0.13$. Figure 1 presents further details.

3.3 Class-based Word Alignment

The proposed alignment algorithm *ClassAlign* is based on the following observations: First, dictionary translations can be used to produce high-precision connections. Thus, *DictAlign* should be employed to produce initial connections whose translations exhibit a high similarity to a DT. That is, a relatively high threshold, 0.7 should be used. Second, the class-based rules acquired through the *ClassRule* Algorithm should capture the diversity of translations to a large extent. According to the observations in Section 2, the rules should stipulate most of the connections left out in the *DictAlign* step. Nevertheless, conflicting connections do occasionally arise. Such conflicts can be resolved according to an additional consideration of distortion mentioned in Section 1

Estimating the Likelihood of a Connection Candidate. The above observations can be stated formally from the perspective of Brown et al.'s (1993) Model 2. As mentioned earlier, the model stipulates that a connection be given a probability value $\Pr(s,t)$, the product of lexical translation probability $\mathrm{t}(s \mid t)$ and distortion probability, $\mathrm{d}(i \mid j,l,m)$. Also according to the model, we give each connection candidate a probabilistic value based on lexical and positional considerations:

$$\Pr(s,t) = \mathrm{t}(s,t) \times \mathrm{d}(i,j) \tag{6}$$

We argue, however, that it is difficult to robustly estimate $\mathrm{t}(s,t)$ and $\mathrm{d}(i,j)$ for all the values of s, t, i, and j. Therefore, the two functions are defined and estimated by a limited number of cases, according to lexical, conceptual, and positional conditions.

For this purpose, we define conceptual similarity between s and t as follows:

$$\text{ConceptSim}(s,t) = \max_{s \in X, t \in Y} \text{ClassSim}(X,Y) \tag{7}$$

Lexical and conceptual conditions are set up based on DTSim and ConceptSim, while positional conditions are set up based on **dislocation**, a distortion measure relative to both left and right context.

Estimation of LTP Based on Lexical and Conceptual Conditions. The LTP $\text{t}(s,t)$ is defined by the following cases:

Case 1. Connection (s,t) exhibits high lexical and conceptual similarity, i.e., ConceptSim $(s,t) \geq h_1$ and DTSim$(s,t) \geq h_2$.
Case 2. Connection (s,t) exhibits high conceptual similarity, i.e., ConceptSim $(s,t) \geq h_1$ and DTSim$(s,t) < h_2$.
Case 3. Connection (s,t) exhibits high lexical similarity, i.e., ConceptSim $(s,t) < h_1$ and DTSim$(s,t) \geq h_2$.
Case 4. Otherwise, ConceptSim $(s,t) < h_1$ and DTSim$(s,t) < h_2$.

The connections satisfying each condition are given the same probability value determined by maximal likelihood estimation (MLE). For instance, if there are k true connections in a sample of n candidates (s,t) such that ConceptSim$(s,t) \geq h_1$ and DTSim$(s,t) \geq h_2$, then all these candidates are given the same MLE value for LTP, i.e., $\text{t}(s,t) = t_1 = k/n$. Equation (8) sums up the above discussion:

$$\text{t}(s,t) = \begin{cases} t_1 \text{ if ConceptSim } (s,t) \geq h_1 \text{ and DtSim}(s,t) \geq h_2, \\ t_2 \text{ if ConceptSim } (s,t) \geq h_1 \text{ and DtSim}(s,t) < h_2, \\ t_3 \text{ if ConceptSim } (s,t) < h_1 \text{ and DtSim}(s,t) \geq h_2, \\ t_4 \text{ if ConceptSim } (s,t) < h_1 \text{ and DtSim}(s,t) < h_2. \end{cases} \tag{8}$$

Estimation of Distortion Probability (DP). In a similar fashion, we formulate the distortion function by cases related to the monotonicity of translational position with respect to context. Such a formulation is inspired by Gale and Church's (1991b) treatment of distortion. In their study, the authors replace the distortion probability with a probability function defined by different values of **slope**, a measure of the position of t with respect to the left context of s. This measure is generally quite accurate, leading to a distribution function concentrating at slope 1. Nevertheless, room for improvement still exists, as can be illustrated using the concept of a binary inversion transduction tree (ITT) proposed by Wu (1995). The ITT is a shared parse tree depicting the structural difference between a sentence S and its translation T. Figure 2 presents the ITT of (E12, C12). The horizontal bar denotes that the noun phrase *such a lazy mortal* and the prepositional phrase *as you* are inverted when translated into Mandarin. The slope of the first word in such an inverted structure is typically quite large, making the distribution of the slope function slightly flat. If multiword structural inversion occurs, as it frequently does, then the slope of the first word according to the right context is still very small.

(E12) I_1 've$_2$ never$_3$ known$_4$ such$_5$ a$_6$ lazy$_7$ mortal$_8$ as$_9$ you$_{10}$.$_{11}$

(C12) 我$_1$ 從來$_2$ 沒有$_3$ 見過$_4$ 像$_5$ 你$_6$ 這麼$_7$ 懶$_8$ 的$_9$ 人$_{10}$ 。$_{11}$

Table 8
Alignment connections for example (E12, C12).

i	0	1	2	3	4	5	6	7		8	9	10	11
s	$	I	've	never	known	such	a	lazy		mortal	as	you	.
t	$	我	從來	沒有	見過	這麼		懶	的	人	像	你	。
j		1	2	3	4	7		8	9	10	5	6	11

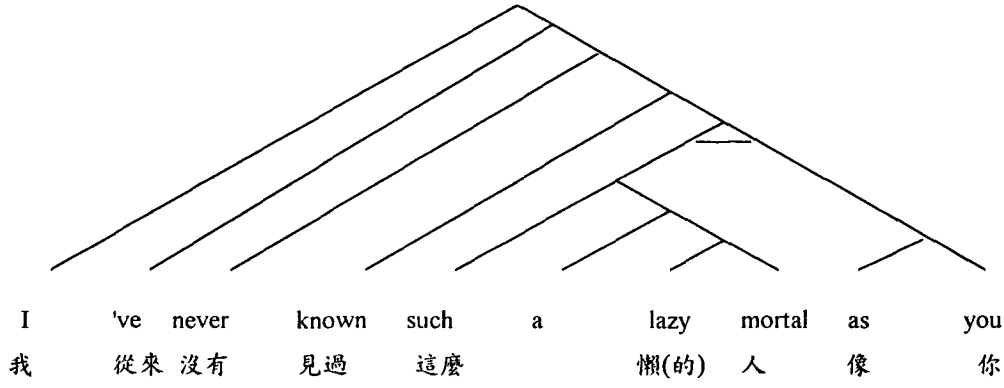

Figure 2
ITT of the example-translation pair (E12, C12).

For instance, the first word *as* in the inverted structure *as you* has a high slope value with respect to its left context *such a lazy mortal*. However, since the words *as* and *you* translate into the fifth and sixth words in (C12), the word *as* has a slope value of 1 with respect to its right context, *you*.

We believe that by considering the translational position relative to both the left and right contexts, one obtains a distribution function with a smaller deviation, thereby making a tighter estimation possible for $d(i,j)$. To this end, we define dislocation, dis, for the connection (s,t) of the ith and jth words in S and T to denote $|(j-j')-(i-i')|$, where i' is the position of a word s' sharing the minimum syntactic structure with s, and s' translates into t', the j'th word in T. Short of syntactic analysis, $\mathrm{dis}(i,j)$ can be calculated with respect to a nearby connection in *CONN*, the initial connections established by *DictAlign*. Such treatment closely approximates the dislocation value. In light of this, dislocation can be defined as follows:

$$\mathrm{dis}(i,j) = \begin{cases} |j-j'| & \text{if } \exists(j')(i,j') \in CONN, \\ \min(|d_L|,|d_R|) & \textit{otherwise}, \end{cases} \tag{9}$$

where
i = the sequence number of s in S,
j = the sequence number of t in T,[4]
$d_L = (j-j_L)-(i-i_L)$,
$d_R = (j-j_R)-(i-i_R)$,
$(i_L,j_L) = \mathrm{argmax}_{(i',j')\in CONN_{<i}}\, i'$,
$(i_R,j_R) = \mathrm{argmin}_{(i',j')\in CONN_{>i}}\, i'$,
$CONN_{<i} = \{(k,l) \mid k\text{th and } l\text{th word in } (S,T) \text{ form a connection in } CONN,\ k<i\}$,

		I	've	never	known	such	a	lazy	mortal	as	you	.
	0	1	2	3	4	5	6	8	10	5	6	11
我	1	**0**	1	2	3	4	5	7	9	4	5	10
從來	2	1	0	1	2	3	4	6	8	3	4	9
沒有	3	2	0	0	1	2	3	5	7	2	3	8
見過	4	3	1	0	0	1	2	4	6	1	2	7
像	5	4	2	1	0	0	1	3	5	0	1	6
你	6	5	3	2	1	0	0	2	4	1	**0**	5
這麼	7	6	4	3	2		0	1	3	2	1	4
懶	8	7	5	4	3	2	1	**0**	2	3	2	3
的	9	8	6	5	4	3	2	1	1	2	3	2
人	10	9	7	6	5	4	3	2	**0**	1	4	1
。	11	10	8	7	6	5	4	3	1	0	5	**0**

Figure 3
Dislocation values for the example-translation pair (E12, C12). Each connection candidate in (E12, C12) is represented as a cell. The connections in *CONN* are shown as a bold face 0. Each of these zero dislocation values extends vertically, incrementing by one for each upward or downward move, resulting in a shaded vertical bar of dislocation values. All other connections take their dislocation values from the minimum diagonal projection of the related cell on the two bounding bars. For instance, the projections of the connection (*such*, 這麼) (shaded in figure) on the left and right bounding bars (*I*-connections and *lazy*-connections) are 2 and 1, respectively. Therefore, the dislocation value is 1, the minimum of 2 and 1.

$$CONN_{>i} = \{(k,l) \mid k\text{th and } l\text{th word in } (S,T) \text{ form a connection in } CONN,\ k > i\},$$
$$CONN = \text{the initial connections established according to DT.}$$

The distortion function defined by cases can now be given according to dislocation values.

$$\mathrm{d}(i,j) = \begin{cases} d_1 \text{ if } \mathrm{dis}(i,j) = 0, \\ d_2 \text{ if } \mathrm{dis}(i,j) = 1, \\ d_3 \text{ if } \mathrm{dis}(i,j) = 2, \\ d_4 \text{ if } \mathrm{dis}(i,j) \geq 3. \end{cases} \tag{10}$$

The connection candidates with small dislocation values tend to be true connections. For instance, 8 out of 15 zero-dislocation connections for (E12, C12) are correct, while only 1 out of 20 candidates with a dislocation of 1 is a true connection. All candidates with dislocation values greater than 1 are false. Figure 3 provides further details. Again, all connections satisfying a certain case in equation (10) are given the same MLE value. For instance, if there are k true connections in a sample of n candidates (i,j) with 0 dislocation, then all of these candidates are given the same MLE value for DP, i.e., $\mathrm{d}(i,j) = d_1 = k/n$ for all i and j such that $\mathrm{dis}(i,j) = 0$.

By using a small sample of 200 sentences from the LecDOCE, the LTP and DP values t_i and d_i for $1 \leq i \leq 4$ can be estimated by the maximum likelihood principle. Tables 9 and 10 summarizes the MLE probabilistic values associated with lexical,

4 The sequence number of Mandarin words is assigned according to the segmentation that satisfies the long-word-first heuristic and is consistent with the established connections in *CONN*.

Table 9
Maximum likelihood estimation (MLE) of LTP.

Conceptual and Lexical Condictions	# Candidates	# True Connections	MLE of t(s,t)	
ConceptSim(s,t) $\geq$ 0.05 and DTSim(s,t) $\geq$ 0.3	508	481	t_1	0.947
ConceptSim(s,t) $\geq$ 0.05 and DTSim(s,t) $<$ 0.3	167	84	t_2	0.503
ConceptSim(s,t) $<$ 0.05 and DTSim(s,t) $\geq$ 0.3	1,589	499	t_3	0.193
ConceptSim(s,t) $<$ 0.05 and DTSim(s,t) $<$ 0.3	14,687	165	t_4	0.011

Table 10
Maximum likelihood estimation (MLE) of DP.

Dislocation	# Candidates	# True Connections	MLE of d(i,j)	
dis = 0	2,158	893	d_1	0.414
dis = 1	3,445	210	d_2	0.061
dis = 2	2,805	31	d_3	0.011
dis $\geq$ 3	9,543	95	d_4	0.010

conceptual, and positional factors. The above description of word alignment is summarized as the *ClassAlign* algorithm.

Algorithm 3. (*ClassAlign*) **Class-based word alignment for a pair of sentences** (S,T).

Step 1: Tag each word in S with POS information and convert each word to the root form to obtain the set W_S of words in S.

Step 2: Initialize the result *ANS* to an empty list. Run *DictAlign* on (S,T) to obtain a list of initial connections, *CONN*.

Step 3: Look up the dictionary to obtain the set W_T of possible words in T.

Step 4: For each connection candidate $(s,t) \in W_S \times W_T$, compute Pr(s,t) according to equations (6) through (10).

Step 5: Add to *ANS* the connection (s^*,t^*) that maximizes Pr(s,t) over all $s,t \in W_S \times W_T$ with a value greater than h_3.[5] Remove all conflicting candidates involving s^* and t^* from subsequent consideration. This step repeats itself until the candidates run out or every remaining candidate (s,t) is associated with a Pr(s,t) value lower than h_3.

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

3.4 An Illustrative Example

In the following, we demonstrate how *ClassAlign* works using example (E10, C10), reproduced below with the sequence number of each word denoted by a subscript number.

(E10) The$_1$ old$_2$ lady$_3$ was$_4$ clad$_5$ in$_6$ a$_7$ fur$_8$ coat$_9$.$_{10}$

(C10) 這$_1$ 位$_2$ 老$_3$ 婦人$_4$ 穿著$_5$ 皮$_6$ 裘$_7$ 。$_8$

As demonstrated earlier in Section 3.1, *DictAlign* produces connections (*old*, 老) and (*clad*, 穿著) from (E10, C10) using a threshold value of 0.7 for *DTSim*. Table 11 lists

5 Ties are resolved in favor of the longer, leftmost Mandarin word.

Table 11
Classes listed in LLOCE for W_S in (E10).

Word	POS	Classes in LLOCE
the	det	Gh285
old	adj	Lg200, Lg208, Lh241
lady	n	Ci157, Ci158, Ci160, Ca005
be	v	Aa001, De080, Li273, Na001, Nf159
clad	adj	Dg136
in	prep	Li272, Mh204
a	det	Nd098
fur	n	Hc088
coat	n	Dg142, Hc093

Table 12
Classes listed in CILIN for W_T in (C10).

Word	POS	Classes in LLOCE
這	det	Ed61
這	n	Ca31, Ka08
位	n	Cb01, Di15, Di17
位	cl	Dn08
老	n	Ab02
老	adj	Eb15, Eb24, Eb29, Eb36, Ec05, Ed51, Ee21
婦人	n	Ab01
婦	n	Ab01, Ah15
人	n	Aa01, Dd17, De01, Dn03
穿著	v	Fa18
皮	n	Bb04, Bc02, Bk10, Bm10, Bm13
皮	adj	Ee09
裘	n	Bq03
裘	cl	Dn08

the W_S words and their relevant POS and topical sets in the LLOCE. Table 12 displays the W_T words and their relevant CILIN categories. Table 13 presents the dislocation values for all connection candidates in $W_S \times W_T$. The cells with a boldface 0 in Table 13 represent the initial connections in *CONN* and two dummy connections placed at the beginning and end of both sentences. Table 14 lists the connection candidates with higher $\Pr(s,t)$ values.

After executing Step 5, *ClassAlign* selects the candidates, (*lady*, 婦人),[6] (*fur*, 皮), both with $\Pr(s,t)$ value of 0.392, in terms at Step 6. These connections are added to *ANS* and the conflicting candidates such as (*old*, 老婦), (*old*, 老婦人), (*old*, 婦人), (*lady*, 婦), (*lady*, 人), (*in*, 皮), (*coat*, 皮), (*fur*, 裘), etc. are removed. In the subsequent iterations, connections (*coat*, 裘) ($\Pr(s,t) = 0.208$), (*old*, 老) ($\Pr(s,t) = 0.080$), (*clad*, 穿著) ($\Pr(s,t) = 0.080$), and (*The*, 這)[7] ($\Pr(s,t) = 0.005$), are selected. *ClassAlign* stops after running out of connections with a probabilistic value greater than h_3, 0.005. Table 15 summarizes the connections chosen to form the solution. The success rate is evaluated

6 This candidate ties with (*lady*, 婦). The conflict is resolved in favor of the longer Mandarin word.
7 This candidate ties with (*The*, 位). The conflict is resolved in favor of the leftmost Mandarin word.

Table 13
Dislocation values for connection candidates (s,t) in (E10, C10).

		The	old	lady	was	clad	in	a	fur	coat	
	0	1	3	3	4	5	4	5	6	7	8
這	1	0	2	2	3	4	3	4	5	6	7
位	2	0	1	1	2	3	2	3	4	5	6
老	3	1	**0**	0	1	2	1	2	3	4	5
婦人	4	2	1	0	0	1	0	1	2	3	4
穿著	5	3	2	1	0	**0**	1	0	1	2	3
皮	6	4	3	2	1	1	0	1	0	1	2
裘	7	5	4	3	2	2	1	0	1	0	1
。	8	6	5	4	3	3	2	1	0	1	**0**

according to how many English words are correctly aligned.[8] 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 a Mandarin word; otherwise it is considered a success. For this example, all nine words are aligned correctly. Therefore, the success rate is 9/9 = 100%.

4. Experiments with *ClassAlign*

To assess 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 a user's manual available in both languages to assess the method's robustness and generality. The similarities and differences between English and Mandarin texts are briefly reviewed, since our experiments involve the alignment of English-Mandarin parallel corpora. A general description of the materials used in the experiments follows. Finally, the success rates are quantitatively evaluated.

4.1 Contrastive Analysis of English and Mandarin Chinese

Language typology is the study of similarities and differences between languages, formalized in terms of parameters such as word order and morphological structure. Li and Thompson (1981) examine Mandarin Chinese according to four typological parameters that reveal the basic structure of Mandarin Chinese as compared to those of other languages, English in particular. These four parameters are the morphological structure of words, the number of syllables per word, topic prominence, and word order. Li and Thompson's typological description of Mandarin is described below, from the perspective of the task of word alignment.

8 A small percentage of connections (7.8%) in our evaluation are incomplete ones and are considered to be correct. Melamed (1996) takes the same stance in his study of deriving a probabilistic lexicon. He observes that even incomplete entries are useful for many applications and there are ways of expanding incomplete morphemes or words in a connection, so that they become complete (Smadja 1992).

Table 14
The connection candidates (s,t) in (E10, C10) with higher Pr(s,t) values.

Connection Candidates				Lexical Translation Probability			Distortion Prob		TP
i	j	s	t	ConceptSim(i,j)	DTSim(i,j)	t(s,t)	dis(i,j)	d(i,j)	Pr (s,t)
1	1	The	這	0	0	0.011	0	0.414	0.005
1	2	The	位	0	0	0.011	0	0.414	0.005
2	3	old	老	0	0.74	0.193	0	0.414	0.080
2	3	old	老婦	0	0.51	0.193	0	0.414	0.080
2	3	old	老婦人	0	0.42	0.193	0	0.414	0.080
3	3	lady	老婦	0	0.30	0.193	0	0.414	0.080
3	4	lady	婦人	0.21	0.39	0.947	0	0.414	0.392
3	3	lady	老婦人	0	0.31	0.193	0	0.414	0.080
3	4	lady	婦	0.21	0.34	0.947	0	0.414	0.392
3	4	lady	人	0	0.56	0.193	0	0.414	0.080
4	3	was	老	0	0	0.011	0	0.414	0.005
4	4	was	婦人	0	0	0.011	0	0.414	0.005
5	5	clad	穿	0	0.71	0.193	0	0.414	0.080
5	5	clad	穿著	0	1.00	0.193	0	0.414	0.080
5	5	clad	著	0	0.62	0.193	0	0.414	0.080
6	6	in	皮	0	0	0.011	0	0.414	0.005
7	7	a	裘	0	0	0.011	0	0.414	0.005
8	6	fur	皮	0.28	0.67	0.947	0	0.414	0.392
9	6	coat	皮	0	0.31	0.193	1	0.061	0.012
9	7	coat	裘	0.14	0	0.503	0	0.414	0.208

Table 15
Final alignment of example (E10, C10). Initial alignment connections are shown in shaded cells.

i	0	1		2	3	4	5	6	7	8	9	10
s	$	The		old	lady	was	clad	in	a	fur	coat	.
t	$	這	位	老	婦人		穿著			皮	裘	。
j	0	1	2	3	4		5			6	7	8

Morphological Structure of Words. The most striking feature of Mandarin as compared to English is the relative simplicity of word structure. That is, most Mandarin words are comprised of a single morpheme rather than a stem morpheme and a suffix serving grammatical functions such as case (as in Turkish and Japanese), number, agreement, or tense (as in many other languages including English). Mandarin verbs do have aspect morphemes, including 了 (*-le*, perfective), 過 (*-guo*, experienced action) and 著(*-zhe*, durative). Other grammatical functions are either non-existent or expressed through an additional function word. In contrast to this lack of inflectional morphological complexity, Mandarin is relatively rich in other types of morphological combinations, including compounding.

These morphological differences result in a difference in the number of words in an English sentence and its Mandarin translation. In terms of alignment, this word-number difference means that multiword connections must be considered, a task which

is beyond the reach of methods proposed in recent alignment works based on Brown et al.'s (1993) Model 1 and 2.

Basic Orientation of the Sentence: Topic vs. Subject. Another feature distinguishing Mandarin from other languages is **topic prominence**. In addition to the grammatical relation of subject, a description of Mandarin must include the topic element, which can be characterized as follows: First, a topic always comes first in the sentence and is optionally followed by a pause in speech. Second, a topic is the old information of which both the speaker and listener have some knowledge. Third, what distinguishes a topic from a subject is that the subject must always have a direct syntactic and semantic relation with the verb, but the topic does not need to. For instance, in the sentence (E13, C13), the first word 大象 (*daxiang*, 'elephant') is the topic and the second word 鼻子 (*bizi*, 'nose') is the subject; 大象 'elephant' is the focus of the discourse, but it is the subject 鼻子 'nose' that is very long; not 大象 'elephant'.

(E13) The elephant has a very long nose.
(C13) 大象 鼻子 很 長。
Daxiang bizi hen chang
Elephant nose very long

The topic prominence of Mandarin sentences represents alignment connections with a large distortion in position, leading to difficulty in estimating the likelihood of a connection according to translational position.

Word Order. Greenberg (1963) stated that the world's languages fall into three word order groups according to the order of the subject (S), verb (V), and object (O) in a simple transitive sentence. A language, in general, belongs to one of three basic word order types, SVO, SOV, and VSO. By this notion, English is an SVO language in which the verb typically follows the subject and precedes the object. For most languages, other aspects of word order, such as that of modifier and modified elements, correlate with the order of V and O. However, Mandarin is not an easy language to classify according to this typology for a number of reasons. First, the notion of subject is not well-defined. Second, unlike in English, word order in Mandarin is not determined solely on grammatical grounds but rather depends on semantics. For instance, whether an adverbial expression appears in pre- or postverbal position depends on subtle semantic differences. More specifically, a time phrase in preverbal position tends to denote punctual time, while that in postverbal position signals durative time, as in:

(E14) I have a meeting at three o'clock.
(C14) 我 三 點鐘 開會。
I three o'clock have-a-meeting
(E15) I slept for three hours.
(C15) 我 睡 了 三個 鐘頭。
I sleep ASPECT three o'clock
(C15′) * 我 三個 鐘頭 睡 了。
I three o'clock sleep ASPECT

In contrast, both kinds of time phrase appear in postverbal position in English. As a result of facts such as these, many linguists contend that Mandarin is a language in transition from SVO to SOV. Further details can be found in Li and Thompson (1981).

Similar to the situation created for topic prominent sentences, the SOV features of Mandarin represent a deviation from the SVO order of English. Such a deviation further worsens our ability to estimate the likelihood of a connection according to translational position.

4.2 The Experimental Setup

The experimental results obtained from the proposed algorithm with respect to word alignment are presented in this section. Nearly 42,000 example sentences and their translations from the LecDOCE were used as training data, primarily to acquire rules and to determine MLE estimates for the cases of LTP and DP. The algorithm's performance was evaluated using the two sets of data. The closed test set consists of 200 examples and their Mandarin translations randomly selected from the LecDOCE. The English examples range from 8 to 23 words long; average example length is 11.5 words. There are, on average, 1.56 inversions per example-translation pair. The open test set consists of 200 sentences randomly drawn from the English and Chinese versions of the LightShip User's Guide. The English sentences in this test set range from 4 to 34 words long; average sentence length is 11.8 words. There are, on average, 1.60 inversions per sentence pair. Table 16 provides some examples from the LightShip User's Guides.

The two thesauri, LLOCE and CILIN, are used as the classification systems of source and target words. The LLOCE contains 23,769 entries and CILIN contains 63,754 entries. Both thesauri cover just over 90% of the words in the test sets.

4.3 Evaluation

The first three experiments were designed to demonstrate the effectiveness of the naive *DictAlign* algorithm based on a bilingual MRD. According to the experimental results, although *DictAlign* produces high-precision alignment, the coverage for both test sets is below 30%. However, if the thesaurus effect is exploited, the coverage can be increased considerably, at the cost of a decrease of less than 4% in precision. Table 17 provides further details.

In the fourth experiment, the *ClassAlign* algorithm is employed to align both sets of test data again. Table 18 reveals that the acquired conceptual information compensates for what is lacking in the LecDOCE to yield optimum alignment results. The *ClassAlign* algorithm expands coverage almost twofold to over 80%, while maintaining the same level of precision. The generality of the approach is evident from the open test's comparably high coverage and precision rates. As shown in Table 18, over 80% of the source words in both test sets are connected to a target and over 90% of the connections are true ones.

5. Discussion

This section thoroughly analyzes the alignment results from the experiments described in Section 4 and, in particular, the data relating to cases where the algorithms failed. Analytical results demonstrate the strengths and limitations of the methods and suggest possible improvements to the algorithms.

5.1 Compounding in Mandarin

As stated earlier, the compounding effect in Mandarin frequently results in a change in the number of words between an English sentence and its Mandarin translation. The correct alignment decision for a Mandarin compound frequently involves more than one English word. *ClassAlign* often fails under such circumstances. For instance,

Table 16
Some examples from LightShip User's Guide.

English Sentence	Mandarin Sentence	Inversion
This chapter introduces the components of applications, screens, and files.	本章將介紹 LightShip 應用、畫面與檔案。	0
Coloring a Screen Background.	為畫面背景著色。	1
You can customize 48 of the colors on the palette by choosing Options Palette.	您可以選擇[Options][Palette] 以更改色盤上的48 種顏色。	2
You can use a customized color to change the color of an object or the screen background color.	您可以自調顏色以改變物件或畫面背景的顏色。	1
The palette is saved with a screen, and not with an object.	色盤是隨著畫面儲存的, 非隨著物件。	1
Assign an attribute from the object's menu.	自選擇表中設定屬性。	1
For information, see Assigning an Attribute to objects on page 3–16.	詳情請見設定物件屬性部份。	1
Selecting an Object in the Debug Window.	選擇 Debug 視窗中的物件。	1
Select the values you want, as shown in Figure 4-11.	2.選出您所要的值(參見圖4-11) 。	1
A user-entry, user-edit document object can store simple variable references.	使用者輸入,使用者編輯文件物件可儲存單一參考變數。	0
If you use Calculate Object when Calculation Manual is on to update the specified objects.	當自動計算物啟時,您可以使用物件計算來更新特定的物件。	3
To select data in a separate source document object:	選擇另一來源文件物件中的資料:	1
4. Define the colors for the low or high value.	4.定義高低值的顏色。	2

Table 17
Experimental results for *DictAlign*.

	Test Set #1: LecDOCE Examples				Test Set #2: LightShip Manual			
	# Matched	# Correct	Coverage	Precision	# Matched	# Correct	Coverage	Precision
DictAlign (DTSim = 1.0)	525	505	28.8%	96.2%	604	576	25.6%	95.4%
DictAlign (DTSim > 0.7)	808	755	44.3%	93.4%	767	705	32.5%	91.9%
DictAlign (DTSim > 0.5)	937	822	51.4%	87.7%	1023	825	43.4%	80.7%

Table 18
Experimental results for *ClassAlign*.

	Test Set #1: LecDOCE Examples				Test Set #2: LightShip Manual			
	# of Words	# Correct	Coverage	Precision	# of Words	# Correct	Coverage	Precisic
All words	1,823	1,460	100%	80.1%	2,359	1,800	100%	76.3%
Matched words	1,561	1,460	85.6%	93.5%	1,965	1,800	83.3%	91.6%

ClassAlign incorrectly connects the compound 劇團 in (C16) to a single English word *company* according to the alignment rule (*Co292*, *Dm07*).

(E16) She is a star with the theatre company.

(C16) 她是劇團的紅星。

Other methods for aligning English and Mandarin texts in the literature also fall prey to the problem of Mandarin compounds. For instance, the following partially correct connections complicated by compounding are reported in a recent study on alignment of Hong Kong Basic Law (Fung and McKeown 1994).

(E17) monoxide

(C17) 一氧化炭 ('carbon monoxide')

(E18) Basic

(C18) 基本法 ('Basic Law')

(E19) second

(C19) 二讀 ('second reading')

Because it is not limited to the connections involved in a presegmented target sentence (Fung and McKeown 1994; Wu and Xai 1994), *ClassAlign* avoids most instances of these errors. In addition, with elaborate preprocessing such as parsing, phrase grouping, and collocation analysis (Smadja 1992), the problem of word-number difference

Table 19
The final alignment of example (E20, C20).

1	2	3	4	5	6	7		8	9	10
He	abdicated	all	responsibility	for	the	care		of	the	child
他	放棄	一切	責任	了		照顧	該	的		小孩
1	2	8	9	3		4	5	7		6

can be averted by performing alignment at various levels: parse tree (Matsumoto, Ishimoto, and Utsuro 1993; Meyers, Yangarber, and Grishman 1996), phrase (Kupiec 1993), and collocation (Smadja, McKeown, and Hatzivassiloglou 1996).

5.2 Function Words, Collocation, and Free Translation

Language-Specific Function Words. The morphological differences between English and Mandarin give rise to many language-specific function words. Such Mandarin function words are often quite ambiguous in part of speech as well as in word sense, leading to numerous alignment errors. For instance, *ClassAlign* connects the words *for* and *of* in (E20) erroneously to the morphemes 了 and 的 in (C20), respectively. Table 19 presents further details.

(E20) He abdicated all responsibility for the care of the child.

(C20) 他放棄了照顧該小孩的一切責任。

Collocation. As mentioned in the previous section, collocation is one of the reasons why in-context translation usually deviates from the dictionary translation. However, unlike other deviations, bilingual collocation is not easily bounded within a couple of classes. For instance, the translation for *take* (*Mb051, carrying, taking and bring*) in the collocation *take effect* is usually 見 ('see') (*Fc04, seeing and looking*), as in example (E21, C21). However, there is insufficient evidence to support a class-to-class mapping from *Mb051* to *Fc04*. In any case, deriving the *Mb051-to-Fc04* mapping would be an overgeneralization.

(E21) How soon does the medicine *take* effect?

(C21) 藥多久見效？

Paraphrased and Free Translations. For various reasons, such as language typology, style, and cultural differences, a translator does not always translate literally on a word-by-word basis. Adding and deleting words is commonplace, sometimes resulting in a paraphrased or free translation. Such translations obviously create problems for word alignment. For instance, in example (E24, C24), only one word, *I*, is translated literally, into 我. The main verb *angle* in example (E25) is given a paraphrased translation 改變角度 ('to change the angle'). The noun phrase *the people she is speaking to* in (E25) is paraphrased as 聽眾 'audience.' A significant amount of free translation arises due to the use of four-morpheme Mandarin idioms for stylistic reasons. For instance, the clause *as long as I breathe* in (E22) translates into an idiom 有生之年 and the sentence

(E23) translates into 入鄉隨俗. Such free or paraphrased translations are beyond the reach of the proposed method.

(E22) I shall love you as long as I breathe!

(C22) 在有生之年我永遠愛你。

(E23) When in Rome, do as the Romans do.

(C23) 入鄉隨俗。

(E24) I don't care who wins.

(C24) 我在這問題上保持中立。

(E25) She angles her reports to suit the people she is speaking to.

(C25) 她改變角度去寫她的報告，以遷就她的聽眾。

5.3 Class-based versus Word-based Models

ClassAlign achieves a degree of generality in the sense that a true connection can be identified, even when it occurs only rarely, or not at all, in the training corpus. This kind of generality is unattainable with statistically trained word-based models. Moreover, class-based models offer the advantages of a smaller storage requirement and higher system efficiency. Unfortunately, they have the disadvantage of erroneous overgeneralization from word-specific connections. For instance, due to the acquired mapping from *Gg273* (element of sound in language) to *Bg07* (sound, tone, etc.), the verb *accent* in (E26) is connected erroneously to 音節 ('syllable') in (C26).

(E26) The accent in the word "important" is on the second syllable.

(C26) Important 這個字的重音是在第二音節。

Nevertheless, our experiment has shown that the advantages outweigh the disadvantages, at least for this particular formulation of a class-based approach to alignment.

6. Concluding Remarks

In this paper, we have presented an algorithm capable of identifying words and their in-context translations in a bilingual corpus. The algorithm is effective for specific linguistic reasons. First, a significant majority of words have diversified translations that are not found in a bilingual dictionary or statistically-derived lexicon but that are largely bounded within the word classes in thesauri. Therefore, we contend that a more successful alignment can be achieved using a class-based approach. Our assumption seems to hold, for the experiments in this study demonstrate that the method provides broad-coverage alignment with almost no loss in precision.

In a broader sense, we have shown that thesauri and corpora can be used in combination to address the critical issues of generality and efficiency. The thesaurus provides classification that can be used to generalize the empirical knowledge gleaned

from a corpus. The corpus provides training and testing materials, thereby allowing knowledge to be derived and evaluated objectively.

The algorithm's performance could definitely be improved by enhancing the various modules of the algorithms, e.g., morphological analyses, bilingual dictionary, monolingual thesauri, and rule acquisition. Nevertheless, this work presents a functional core for processing bilingual corpora at lexical and conceptual levels.

While this paper has specifically addressed English-Chinese corpora, the linguistic issues motivating the algorithms seem to be quite general and are, to a large extent, language independent, which means that the algorithm presented here should be adaptable to other language pairs. The prospects for English-Japanese or Chinese-Japanese, in particular, seem highly promising.

Acknowledgments

This work was partially supported by ROC NSC grants 82-0408-E-007-195, 83-0408-E-007-010, 84-2213-E-007-023. We are grateful to Keh-Yih Su for his suggestions and comments at the early stage in the development of this work. We are also thankful to J-P Chanod and the anonymous reviewers for many useful suggestions. We would like to thank Liming Yu from Zebra English Service Union, Betty Teng and Nora Liu from Longman Asia Limited, Keh-Jiann Chen and Chu-Ren Huang from Academy Sinica, and Perry Chang from Galaxy Software Services for making the dictionaries, thesauri, and corpora available to us.

References

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.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, J. D. Lafferty, and R. L. Mercer. 1992. Analysis, statistical transfer, and synthesis in machine translation. In *Proceedings of the Fourth International Conference on Theoretic and Methodological Issues in Machine Translation*, pages 83–100.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Brown, P. F., V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

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

Chang, J. S. and M. H. C. Chen. 1994. Using partial aligned parallel text and part-of-speech information in word alignment. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 16–23, Columbia, MD.

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 (AMTA)*, pages 115–124, Montreal, Canada.

Chapman, R. 1977. *Roget's International Thesaurus*. Harper and Row, New York.

Chen, Stanley F. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Meeting*, pages 9–16, Ohio. Association for Computational Linguistics.

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*, pages 1–12.

Church, K. W. and W. A. Gale. 1991. Concordances for parallel text. In *Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research*, pages 40–62, St. Catherine's College, Oxford, England.

Dagan, I., 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*, pages 1–8, Columbus, OH.

Daille, B., E. Gaussier, and J.-M. Lange.

1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 515–521, Kyoto, Japan.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Journal of Ecology*, 26:297–302.

van der Eijk, P. 1993. Automating the acquisition of bilingual terminology. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*, pages 113–119, Utrecht, The Netherlands.

Fujii, H. 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*, pages 237–246.

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 (AMTA-94)*, pages 81–88, Columbia, MD.

Galaxy Software Services. 1994. *Lightship User's Guide* (in Chinese). Galaxy Software Services, Taiwan.

Gale, W. A. and K. W. Church. 1991a. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting*, pages 177–184, Berkeley, CA. Association for Computational Linguistics.

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*, pages 152–157, Pacific Grove, CA.

Gale, W. A. and K. W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Gale, W. A., K. W. Church, and D. 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*, pages 101–112, Montreal, Canada.

Greenberg, J. H. 1963. *Universals of Language*. MIT Press, Cambridge, MA.

Isabelle, P. 1992. Bi-textual aids for translators. In *Proceedings of the Eighth Annual Conference of the UW Centre for the New OED and Text Research*, pages 76–89, Waterloo, Canada.

Kay, M. and M. Röscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

Klavans, J. L. and E. Tzoukermann. 1990. The BICORD system. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 174–179, Helsinki, Finland.

Kupiec, J. M. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting*, pages 17–22, Columbus, OH. Association for Computational Linguistics.

Li, C. N. and S. A. Thompson. 1981. *Mandarin Chinese—A Functional Reference Grammar*. University of California Press, Los Angeles, CA.

Li, Hung-Wen. 1994. Word Alignment and Refinement of Transfer Dictionary. Master thesis, Institute of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, R.O.C.

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

Macklovitch, E. 1994. Using bi-textual alignment for translation validation: The TransCheck system. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 157–168, Columbia, MD.

Matsumoto, Y., H. Ishimoto, and T. Utsuro. 1993. Structural matching of parallel texts. In *Proceedings of the 31st Annual Meeting*, pages 22–30, Columbus, OH. Association for Computational Linguistics.

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

McRoy, Susan W. 1992. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30.

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

Melamed, I. Dan. 1996. Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 125–134, Montreal, Canada.

Meyers, A., R. Yangarber, and R. Grishman. 1996. Alignment of shared forests for bilingual corpora. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 460–465, Copenhagen, Denmark. COLING-96.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to Wordnet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244.

Pilot Software Inc. 1993. *LightShip User's Guide*, Pilot Software Inc., Boston.

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

Shemtov, H. 1993. Text alignment in a tool for translating revised documents. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*, pages 449–453, Utrecht, The Netherlands.

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 (TMI-92)*, pages 67–81, Montreal, Canada.

Smadja, F. 1992. How to compile a bilingual collocation lexicon automatically. In *Proceedings of the AAAI-92 Workshop on Statistically-Based NLP Techniques*, pages 65–71, San Jose, CA. American Association for Artificial Intelligence.

Smadja, F., K. R. McKeown, and V. Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

des Tombe, L. and S. Armstrong-Warwick. 1993. Using function words to measure translation quality. In *Proceedings of the Ninth Annual Conference of the UW Centre for the New OED and Text Research*, pages 1–17.

Wu, D. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Meeting*, pages 80–87, Las Cruces, NM. Association for Computational Linguistics.

Wu, D. 1995. Grammarless extraction of phrasal translation examples from parallel texts. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 354–372, Belgium.

Wu, D. and X. Xia. 1994. Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 206–213, Columbia, MD.