Machine Translation, 9:3-4, 285–313 (1995)

Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon

DEKAI WU dekai@cs.ust.hk
XUANYIN XIA samxia@cs.ust.hk
Department of Computer Science, HKUST,
University of Science & Technology, Clear Water Bay, Hong Kong

Received September 1, 1994; Revised June 1, 1995

Abstract. We report experimental results on automatic extraction of an English-Chinese translation lexicon, by statistical analysis of a large parallel corpus, using limited amounts of linguistic knowledge. To our knowledge, these are the first empirical results of the kind between an Indo-European and non-Indo-European language for any significant vocabulary and corpus size. The learned vocabulary size is about 6,500 English words, achieving translation precision in the 86–96% range, with alignment proceeding at paragraph, sentence, and word levels. Specifically, we report (1) progress on the HKUST English-Chinese Parallel Bilingual Corpus, (2) experiments supporting the usefulness of restricted lexical cues for statistical paragraph and sentence alignment, and (3) experiments that question the role of hand-derived monolingual lexicons for automatic word translation acquisition. Using a hand-derived monolingual lexicon, the learned translation lexicon averages 2.33 Chinese translations per English entry, with a manually-filtered precision of 95.1%, and an automatically-filtered weighted precision of 86.0%. We then introduce a fully automatic two-stage statistical methodology that is able to learn translations for collocations. A statistically-learned monolingual Chinese lexicon is first used to segment the Chinese text, before applying bilingual training to produce 6,429 English entries with 2.25 Chinese translations per entry. This method improves the manually-filtered precision to 96.0% and the automatically-filtered weighted precision to 91.0%, an error rate reduction of 35.7% from using a hand-derived monolingual lexicon.

Keywords: lexical acquisition, translation lexicon, parallel corpus, statistical and corpus-based NLP, English-Chinese machine translation

1. Introduction

A criticism of statistical machine translation tools is that convincing empirical results to date are largely confined to similar language pairs, such as French and English. We offer some contributions to the pool of evidence supporting the language-independence of statistical techniques. Specifically, we report accuracy rates for a new performance measure: the precision of statistical translation lexicon acquisition, between English and Chinese.

The SILC project at HKUST is studying machine learning of natural language translation. In the first phase of SILC (*s*tatistical *i*nter-*l*ingual *c*onversion), as reported herein, we have (1) collected a bilingual corpus of parallel English and Chinese text, (2) aligned the paragraphs and sentences within the corpus, and (3) learned a bilingual lexicon from the aligned data.

One motivation for this work is to conduct English-Chinese "acid tests" of statistical NLP techniques. Another benefit of the approach is that it obtains not only a translation lexicon, but also the *probabilities* of alternative translations for the same word. A further advantage of the learning approach is the ability to acquire lexicons that are adapted for particular domains or genres. The vocabulary of our corpus, for example, includes a high proportion of words not found in the English-Chinese machine-readable dictionaries we have seen.

A recurrent theme of the project is to evaluate the utility of *a priori* linguistic knowledge, relative to self-organizing statistical methods. Specifically, in this work we examine the issue with respect to the tasks of sentence alignment and word translation learning. For sentence alignment, a small, restricted set of lexical cues turn out to raise performance significantly from a purely-statistical baseline model. But for word translation learning, a manually-encoded monolingual lexicon turns out to actually *hamper* performance when compared against using a statistically-derived one.

This article charts the progress of the project's translation lexicon extraction, discussing the design issues and our empirical resolution of the various tradeoffs. We begin in Section 2 with a description of the corpus itself. Section 3 introduces the statistical paradigm within which paragraphs and sentences are aligned. The applicability of a well-known purely statistical method (Gale and Church, 1991) to English-Chinese, an etymologically-unrelated language pair, is empirically tested in Section 4. Though performance turns out to be surprisingly good, a further extension of the method in Section 5 to incorporate limited lexical criteria is shown to improve accuracy significantly. The aligned sentence pairs are fed into the word translation training procedure in Section 6. Section 7 discusses a method of choosing the correct translations for a single word that has multiple possible translations. Finally, in Section 8 we show that automatic *monolingual* lexical acquisition of Chinese further improves the precision of bilingual lexical acquisition to a surprising extent.

2. The English-Chinese Corpus

Though large parallel bilingual corpora are relatively scarce compared with monolingual corpora, they have generated highly interesting results that cannot be obtained using monolingual corpora. Significant progress has been made on problems including automatic sentence alignment (Kay and Röscheisen, 1988; Catizone, Russell, and Warwick, 1989; Gale and Church, 1991; Brown, Lai, and Mercer, 1991; Chen, 1993), coarse alignment (Church, 1993), statistical machine translation (Brown et al., 1990; Brown et al., 1993), word alignment (Dagan, Church, and Gale, 1993), word sense disambiguation (Gale, Church, and Yarowsky, 1993), and collocation learning (Smadja and McKeown, 1994), all exploiting parallel corpora. The dearth of work on non-Indo-European languages can partly be attributed to a lack of the requisite bilingual corpora. To facilitate empirical studies that cannot rely on shared characteristics of Indo-European languages, we have been constructing the HKUST English-Chinese Parallel Bilingual Corpus. To be included, materials must contain primarily tight, literal sentence translations. This rules out most fiction and literary material.

We have been concentrating on bilingual parliamentary proceedings of the Hong Kong Legislative Council (LegCo). Analogously to the bilingual texts of the Canadian Hansard (Gale and Church, 1991), LegCo transcripts are kept in full translation in both English and Cantonese.[1] However, unlike the Canadian Hansard, the original materials were not designed to be available in machine-readable form. We have obtained these materials by arrangement with governmental authorities; their obscure format has necessitated heavy conversion and reformatting, using both manual and automatic processing.

The materials contain high-quality literal translation. Statements in LegCo may be made using either English or Cantonese, and are transcribed in the original language. A translation to the other language is made later to yield complete parallel texts, with annotations specifying the source language used by each speaker. Most sentences are translated 1-for-1. A small proportion are 1-for-2 or 2-for-2, and on rare occasion 1-for-3, 3-for-3, or other configurations.

The emphasis is on clean text so that markup is minimal, but when needed TEI conformant SGML annotation is used (Sperberg-McQueen and Burnard, 1992). Each session occupies a single English file and a single Chinese file. Aside from the standard `<body>` and `</body>` text delimiters for each file, the only other markup at present are the paragraph delimiters `<p>` and `</p>`, and the segment delimiters `<s>` and `</s>` which are used to mark sentences. We use the term "sentence" in a generalized sense including lines in itemized lists, headings, and other nonsentential segments smaller than a paragraph. Samples of the English and Chinese texts are shown in Figures 1 and 2.

Because of the obscure format of the original data, it has been necessary to employ a substantial amount of automatic conversion and reformatting. Sentences are identified automatically using heuristics that depend on punctuation and spacing. Sentence and paragraph segmentation errors occur occasionally, due either to typographical errors in the original data, or to inadequacies of our automatic conversion heuristics. This simply results in incorrectly placed delimiters; it does not remove any text from the corpus.

For the experiments reported in this paper, we used a portion of the corpus occupying approximately 29Mb of raw English text and 15.5Mb of corresponding raw Chinese translation. The English text included nearly 5 million English words (Chinese words are hard to count, as discussed below).

3. Statistical Paragraph and Sentence Alignment

Recently, a number of automatic techniques for aligning sentences in parallel bilingual corpora have been proposed (Kay and Röscheisen, 1988; Catizone, Russell, and Warwick, 1989; Gale and Church, 1991; Brown, Lai, and Mercer, 1991). Such corpora contain the same material that has been translated by human experts into

```
<p>
<s>
DR TANG SIU - TONG ( in Cantonese ) :
</s>
<s>
The question I am going to ask involves a real issue, not a
 hypothetical one.
</s>
<s>
The problem of inflation, which is a major worry in Hong Kong at
 present, has not been discussed in detail in your Policy Address.
</s>
<s>
Mr Governor, what solutions do you have in combating inflation?
</s>
</p>
<p>
<s>
THE GOVERNOR:
</s>
<s>
Maiden questions are not supposed to be that difficult.
</s>
<s>
Inflation in Hong Kong -- and I do not want to sound too much
 like an economist -- is largely a structural consequence of our
 circumstances, largely a consequence of the fact that we are
 growing very fast, but growing very fast with considerable
 restraints because of our geographical circumstances.
</s>
<s>
We have problems of land supply; we have problems of labour
 supply, and those supply side problems at a time when we are
 expanding as rapidly as we are have helped to produce inflation.
</s>
</p>
```

Figure 1. A sample of English text.

<p>
<s>
鄧兆棠議員問:
</s>
<s>
我要問的並不是一個假設性的問題，而是一個真實的問題。
</s>
<s>
施政報告內，沒有詳細講及通脹問題，而通脹卻是現時香港的一大隱憂。
</s>
<s>
請問總督先生，有甚麼辦法可以解決通脹?
</s>
</p>
<p>
<s>
總督答(譯文):
</s>
<s>
這是你第一次提出問題，不應太難解答。
</s>
<s>
香港的通脹－我不希望我的口吻太過像經濟學家－主要是經濟結構所致;
主要是我們的經濟增長很快，但這個快速的增長卻因地理環境而受到相當大的掣肘，
以致出現這個後果。
</s>
<s>
我們有土地不足的問題，我們有勞工短缺的問題。
</s>
<s>
而這些資源不足的問題，在我們的經濟以目前這樣快的速度增長下，便會引致通脹。
</s>
</p>

Figure 2. A sample of Chinese text.

two languages. The goal of alignment is to identify matching sentences between the languages. Alignment is the first stage in extracting structural information and statistical parameters from bilingual corpora. The problem is made more difficult because a sentence in one language may correspond to multiple sentences in the other; worse yet, sometimes several sentences' content is distributed across multiple translated sentences.

Approaches to alignment fall into two main classes: lexical and statistical. Lexically based techniques use extensive online bilingual lexicons to match sentences. In contrast, statistical techniques require almost no prior knowledge and are based solely on the lengths of sentences. The empirical results to date suggest that statistical methods yield performance superior to that of currently available lexical techniques.

However, as far as we know, all work on automatic alignment has been restricted to Indo-European languages. This methodological flaw weakens the arguments in favor of either approach, since it is unclear to what extent a technique's superiority depends on the similarity between related languages. The work reported herein moves towards addressing this problem.

The statistical approach to alignment can be summarized as follows: choose the alignment that maximizes the probability over all possible alignments, given a pair of parallel texts. Formally, choose

$$\arg\max_{\mathcal{A}} \Pr(\mathcal{A}|\mathcal{T}_1, \mathcal{T}_2) \tag{1}$$

where $\mathcal{A}$ is an alignment, and $\mathcal{T}_1$ and $\mathcal{T}_2$ are the English and Chinese texts, respectively. An alignment $\mathcal{A}$ is a set consisting of $L_1 \rightleftharpoons L_2$ pairs where each L_1 or L_2 is an English or Chinese passage.

This formulation is so extremely general that it is difficult to argue against its pure form. More controversial are the approximations that must be made to obtain a tractable version.

The first commonly made approximation is that the probabilities of the individual aligned pairs within an alignment are independent, i.e.,

$$\Pr(\mathcal{A}|\mathcal{T}_1, \mathcal{T}_2) \approx \prod_{(L_1 \rightleftharpoons L_2) \in \mathcal{A}} \Pr(L_1 \rightleftharpoons L_2|\mathcal{T}_1, \mathcal{T}_2) \tag{2}$$

The other common approximation is that each $\Pr(L_1 \rightleftharpoons L_2|\mathcal{T}_1, \mathcal{T}_2)$ depends not on the entire texts, but only on the contents of the specific passages within the alignment:

$$\Pr(\mathcal{A}|\mathcal{T}_1, \mathcal{T}_2) \approx \prod_{(L_1 \rightleftharpoons L_2) \in \mathcal{A}} \Pr(L_1 \rightleftharpoons L_2|L_1, L_2) \tag{3}$$

Maximization of this approximation to the alignment probabilities is easily converted into a minimum-sum problem:

$$\arg\max_{\mathcal{A}} \Pr(\mathcal{A}|\mathcal{T}_1, \mathcal{T}_2) \approx \arg\max_{\mathcal{A}} \prod_{(L_1 \rightleftharpoons L_2) \in \mathcal{A}} \Pr(L_1 \rightleftharpoons L_2|L_1, L_2)$$

$$= \arg\min_{\mathcal{A}} \sum_{(L_1 \rightleftharpoons L_2) \in \mathcal{A}} -\log \Pr(L_1 \rightleftharpoons L_2 | L_1, L_2) \qquad (4)$$

The minimization can be implemented using a dynamic programming strategy.

Further approximations vary according to the specific method being used. Below, we first discuss a pure length-based approximation, then a method with lexical extensions. Section 4 reports experiments addressing the applicability of a suitably modified version of Gale and Church's (1991) length-based statistical method to the task of aligning English with Chinese. Following, Section 5, describes an improved statistical method that also permits domain-specific lexical cues to be incorporated probabilistically.

4. Applicability of Length-Based Methods

Length-based alignment methods are based on the following approximation to equation 4:

$$\Pr(L_1 \rightleftharpoons L_2 | L_1, L_2) \approx \Pr(L_1 \rightleftharpoons L_2 | l_1, l_2) \qquad (5)$$

where $l_1 = \text{length}(L_1)$ and $l_2 = \text{length}(L_2)$, measured in number of characters. In other words, the only feature of L_1 and L_2 that affects their alignment probability is their length. Note that there are other length-based alignment methods that measure length in number of words instead of characters (Brown, Lai, and Mercer, 1991). However, since Chinese text consists of an unsegmented character stream without marked word boundaries, it would not be possible to count the number of words in a sentence without first parsing it.

Although it has been suggested that length-based methods are language independent (Gale and Church, 1991; Brown, Lai, and Mercer, 1991), they may in fact rely to some extent on length correlations arising from the historical relationships of the languages being aligned. If translated sentences share cognates, then the character lengths of those cognates are of course correlated. Grammatical similarities between related languages may also produce correlations in sentence lengths.

Moreover, the combinatorics of non-Indo-European languages can depart greatly from Indo-European languages. In Chinese, the majority of words are just one or two characters long (though collocations up to four characters are also common). At the same time, there are several thousand characters in daily use, as in conversation or newspaper text. Such lexical differences make it even less obvious whether pure sentence-length criteria are adequately discriminating for statistical alignment.

Our first goal, therefore, is to test whether purely length-based alignment results can be replicated for English and Chinese, languages from unrelated families. However, before length-based methods can be applied to Chinese, it is first necessary to generalize the notion of "number of characters" to Chinese strings, because most Chinese text (including our corpus) includes occasional English proper names and abbreviations, as well as punctuation marks. Our approach is to count each Chinese character as having length 2, and each English or punctuation character as

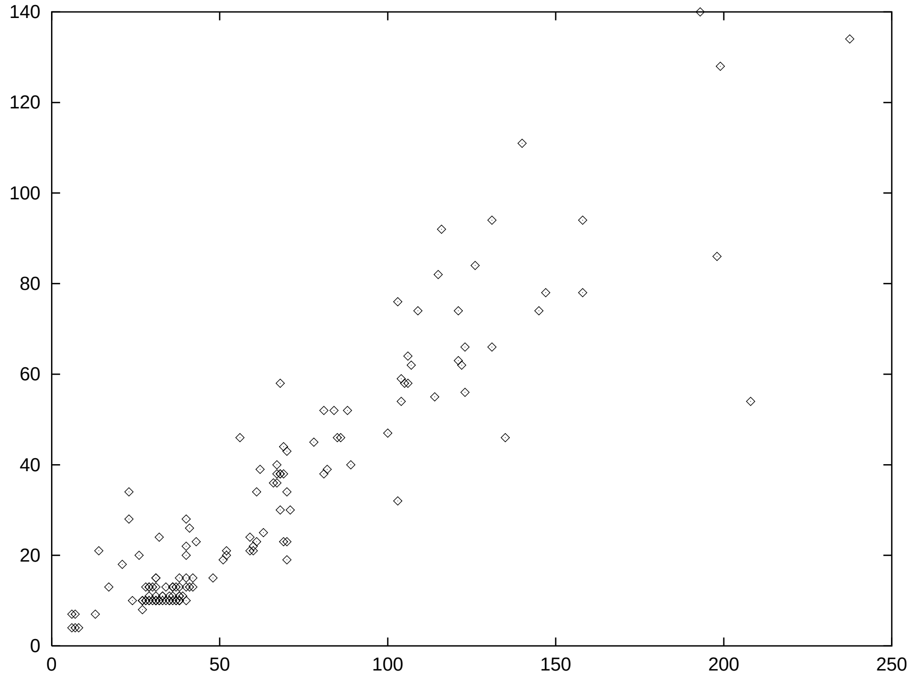

Figure 3. English versus Chinese sentence lengths.

having length 1. This corresponds to the byte count for text stored in the hybrid English-Chinese encoding system known as *Big 5*.

Gale and Church's (1991) length-based alignment method is based on the model that each English character in L_1 is responsible for generating some number of characters in L_2. This model leads to a further approximation which encapsulates the dependence to a single parameter δ that is a function of l_1 and l_2:

$$\Pr(L_1 \rightleftharpoons L_2 | L_1, L_2) \approx \Pr(L_1 \rightleftharpoons L_2 | \delta(l_1, l_2)) \tag{6}$$

However, it is much easier to estimate the distributions for the inverted form obtained by applying Bayes' Rule:

$$\Pr(L_1 \rightleftharpoons L_2 | \delta) = \frac{\Pr(\delta | L_1 \rightleftharpoons L_2)\Pr(L_1 \rightleftharpoons L_2)}{\Pr(\delta)} \tag{7}$$

where $Pr(\delta)$ is a normalizing constant that can be ignored during minimization. The other two distributions are estimated as follows.

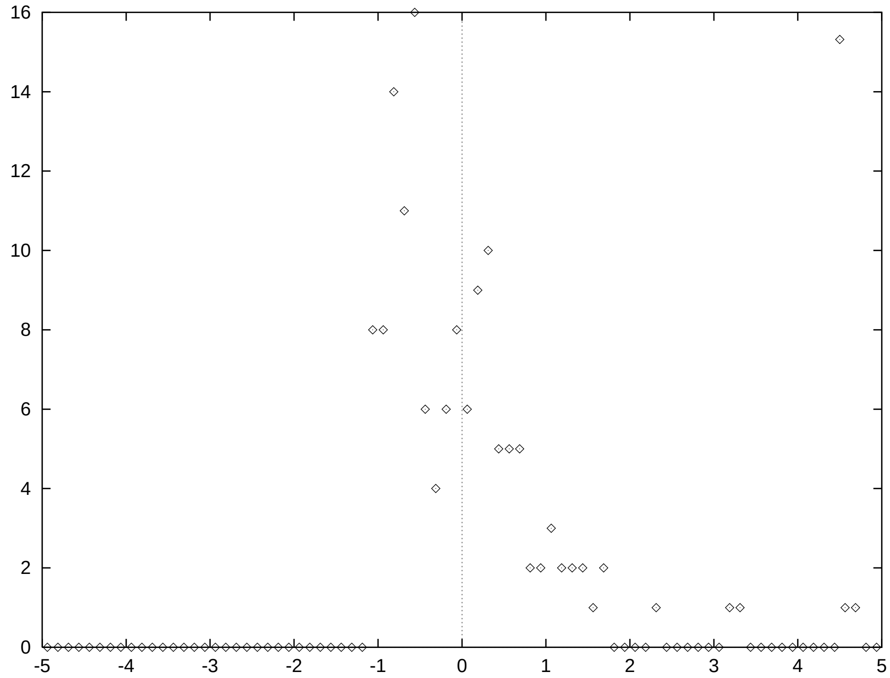

Figure 4. Normalized difference of English and Chinese sentence lengths.

First we choose a function for $\delta(l_1, l_2)$. To do this we look at the relation between l_1 and l_2 under the generative model. Figure 3 shows a plot of English versus Chinese sentence lengths for a hand-aligned sample of 142 sentences. If the sentence lengths were perfectly correlated, the points would lie on a diagonal through the origin. We estimate the slope of this idealized diagonal $c = E(r) = E(l_2/l_1)$ by averaging over the training corpus of hand-aligned $L_1 \rightleftharpoons L_2$ pairs, weighting by the length of L_1. In fact this plot displays substantially greater scatter than the English-French data of Gale and Church (1991).[2] The mean number of Chinese characters generated by each English character is $c = 0.506$, with a standard deviation $\sigma = 0.166$.

We now assume that $l_2 - l_1 c$ is normally distributed, following Gale and Church, and transform it into a new gaussian variable of standard form (i.e., with mean 0 and variance 1) by appropriate normalization:

$$\frac{l_2 - l_1 c}{\sqrt{l_1 \sigma^2}} \tag{8}$$

This is the quantity that we choose to define as $\delta(l_1, l_2)$. Consequently, for any two pairs in a proposed alignment, $\Pr(\delta | L_1 \rightleftharpoons L_2)$ can be estimated according to the gaussian assumption.

To check how accurate the gaussian assumption is, we can use equation 8 to transform the same training points from Figure 3 and produce a histogram. The result is shown in Figure 4. Again, the distribution deviates from a gaussian distribution substantially more than Gale and Church report for French/German/English. Moreover, the distribution does not resemble any smooth distribution at all, including the logarithmic normal used by Brown, Lai, and Mercer (1991), raising doubts about the potential performance of pure length-based alignment.

Continuing nevertheless, to estimate the other term $\Pr(L_1 \rightleftharpoons L_2)$, a prior over six classes is constructed, where the classes are defined by the number of passages included within L_1 and L_2. Table 1 shows the probabilities used. These probabilities are taken directly from Gale and Church; slightly improved performance might be obtained by estimating these probabilities from our corpus.

Table 1. Priors for $\Pr(L_1 \rightleftharpoons L_2)$.

# segments		
L_1	L_2	$\Pr(L_1 \rightleftharpoons L_2)$
0	1	0.0099
1	0	0.0099
1	1	0.89
1	2	0.089
2	1	0.089
2	2	0.011

The aligned results using this model were evaluated by hand for the entire contents of a randomly selected pair of English and Chinese files corresponding to a complete session, comprising 506 English sentences and 505 Chinese sentences. Figures 5 and 6 show an excerpt from this output. Most of the true 1-for-1 pairs are aligned correctly. In Figure 5(4), two English sentences are correctly aligned with a single Chinese sentence. However, the English sentences in (6, 7) are incorrectly aligned

1.	¶MR FRED LI (in Cantonese) : ⌋	¶李華明議員問: ⌋
2.	I would like to talk about public assistance. ⌋	我想談及公共援助問題。 ⌋
3.	I notice from your address that under the Public Assistance Scheme, the basic rate of \$825 a month for a single adult will be increased by 15% to \$950 a month. ⌋	施政報告提到提高單身人士的公共援助基本金額, 由每月825元提高至950元, 即加幅是15%。 ⌋
4.	However, do you know that the revised rate plus all other grants will give each recipient no more than \$2000 a month? On average, each recipient will receive \$1600 to \$1700 a month. ⌋	但你知否經過調整後, 即使加上所有其他津貼, 每名受助者每月所得到的公共援助都不會超過2000元, 平均來說, 他們每月所得的是1600元至1700元左右。 ⌋
5.	In view of Hong Kong's prosperity and high living cost, this figure is very ironical. ⌋	以香港的繁榮和生活水平之高, 這數字根本是一個很大的諷刺。 ⌋
6.	May I have your views and that of the Government? ⌋	請問政府或總督先生, 你有何看法, 是否覺得應全面檢討公共援助的計算方式? ⌋
7.	Do you think that a comprehensive review should be conducted on the method of calculating public assistance? ⌋	因為基數這麼低, 就算加多20%至30%, 仍是遠遠落後於現時的生活水平。 ⌋
8.	Since the basic rate is so low, it will still be far below the current level of living even if it is further increased by 20% to 30%. If no comprehensive review is carried out in this aspect, this " safety net " cannot provide any assistance at all for those who are really in need. ⌋	若不全面檢討公共援助的計算方法, 這安全網根本不能為那些真正有需要的人士提供協助。 ⌋
9.	I hope Mr Governor will give this question a serious response. ⌋	希望總督先生認真回應這問題。 ⌋

Figure 5. A sample of length-based alignment output.

1.	¶THE GOVERNOR: ⌋	¶總督答(譯文): ⌋
2.	It is not in any way to belittle the importance of the point that the Honourable Member has made to say that, when at the outset of our discussions I said that I did not think that the Government would be regarded for long as having been extravagant yesterday, I did not realize that the criticisms would begin quite as rapidly as they have. ⌋	我在昨天的討論開始時說, 我相信政府不會長期被指為揮霍無度。 當時我沒有料到批評會來得這麼快。 ⌋
3.	The proposals that we make on public assistance, both the increase in scale rates, and the relaxation of the absence rule, are substantial steps forward in Hong Kong which will, I think, be very widely welcomed. ⌋	我說這句話, 絕對無意貶低這位議員剛才所提意見的重要性。 我們就公共援助提出的建議, 不論是增加援助金額或是放寬離港期限的規定, 對本港來說, 可說是向前跨進一大步, 我相信普遍會受到歡迎。 ⌋
4.	But I know that there will always be those who, I am sure for very good reason, will say you should have gone further, you should have done more. ⌋	不過, 我知道有些人一定會說, 你應更向前邁進一步, 你應該做多一些, 我肯定他們這樣說是有非常充分的理由。 ⌋
5.	Societies customarily make advances in social welfare because there are members of the community who develop that sort of case very often with eloquence and verve. ⌋	很多社會慣於改善其社會福利, 原因是有些人經常利用動聽的說話及淩厲的詞鋒, 提出這方面的意見。 ⌋

Figure 6. A sample of length-based alignment output (cont'd).

1-for-1 instead of 2-for-1. Also, Figure 6(2, 3) shows an example of a 3-for-1, 1-for-1 sequence that the model has no choice but to align as 2-for-2, 2-for-2.

Judging relative to a manual alignment of the English and Chinese files, a total of 86.4% of the true $L_1 \rightleftharpoons L_2$ pairs were correctly identified by the length-based method. However, many of the errors occurred within the introductory session header, whose format is domain-specific (discussed below). If the introduction is discarded, then the proportion of correctly aligned pairs rises to 95.2%, a respectable rate especially in view of the drastic inaccuracies in the distributions assumed. A detailed breakdown of the results is shown in Table 2. For reference, results reported for English/French generally fall between 96% and 98%. However, all of these numbers should be interpreted as highly domain dependent, with very small sample size.

The above rates are for Type I errors. The alternative measure of accuracy on Type II errors is useful for machine translation applications, where the objective is to extract only 1-for-1 sentence pairs, and to discard all others. In this case, we are interested in the proportion of 1-for-1 *output* pairs that are true 1-for-1 pairs. (In information retrieval terminology, this measures precision whereas the above measures recall.) In the test session, 438 1-for-1 pairs were output, of which 377, or 86.1%, were true matches. Again, however, by discarding the introduction, the accuracy rises to a surprising 96.3%.

The introductory session header exemplifies a weakness of the pure length-based strategy, namely, its susceptibility to long stretches of passages with roughly similar lengths. In our data this arises from the list of council members present and absent at each session (Figure 7), but similar stretches can arise in many other domains. In such a situation, two slight perturbations may cause the entire stretch of passages between the perturbations to be misaligned. These perturbations can easily arise from a number of causes, including slight omissions or mismatches in the original parallel texts, a 1-for-2 translation pair preceding or following the stretch of passages, or errors in the heuristic segmentation preprocessing. Substantial penalties may occur at the beginning and ending boundaries of the misaligned region, where the perturbations lie, but the misalignment between those boundaries incurs little penalty, because the mismatched passages have apparently matching lengths. This problem is apparently exacerbated by the non-alphabetic nature of Chinese. Because Chinese text contains fewer characters, character length is a less discriminating feature, varying over a range of fewer possible discrete values than the corresponding English. The next section discusses a solution to this problem.

Table 2. Detailed breakdown of length-based alignment results.

	1-1	1-2	2-1	2-2	1-3	3-1	3-3
Total	433	20	21	2	1	1	1
Correct	361	17	20	0	0	0	0
Incorrect	11	3	1	2	1	1	1
% Correct	87.1	85.0	95.2	0.0	0.0	0.0	0.0

1.	1 HONG KONG LEGISLATIVE COUNCIL – 8 October 1992 ⌋	立法局會議過程正式紀錄 ⌋
2.	¶HONG KONG LEGISLATIVE COUNCIL – 8 October 1992 1 ⌋	¶一九九二年十月八日星期四 ⌋
3.	¶OFFICIAL RECORD OF PROCEEDINGS ⌋	¶下午二時三十分開始會議 ⌋
4.	¶Thursday, 8 October 1992 ⌋	¶出席者: ⌋
5.	¶The Council met at half - past Two o'clock PRESENT ⌋	¶總督彭定康先生(主席) ⌋
6.	¶THE PRESIDENT HIS EXCELLENCY THE GOVERNOR THE RIGHT HONOURABLE CHRISTOPHER FRANCIS PATTEN ⌋	¶副主席施偉賢議員, C.B.E., Q.C., J.P. ⌋
7.	¶THE DEPUTY PRESIDENT THE HONOURABLE JOHN JOSEPH SWAINE, C.B.E., Q.C., J.P. ⌋	¶布政司霍德爵士議員, K.B.E., L.V.O., J.P. ⌋
8.	¶THE CHIEF SECRETARY THE HONOURABLE SIR DAVID ROBERT FORD, K.B.E., L.V.O., J.P. ⌋	¶財政司麥高樂議員, C.B.E., J.P. ⌋
9.	¶THE FINANCIAL SECRETARY THE HONOURABLE NATHANIEL WILLIAM HAMISH MACLEOD, C.B.E., J.P. ⌋	¶律政司馬富善議員, C.M.G., J.P. ⌋
10.	(37 misaligned matchings omitted)	
11.	¶THE HONOURABLE MAN SAI - CHEONG ⌋	¶潘國濂議員 ⌋
12.	¶THE HONOURABLE STEVEN POON KWOK - LIM THE HONOURABLE HENRY TANG YING - YEN, J.P. ⌋	¶唐英年議員, J.P. ⌋
13.	¶THE HONOURABLE TIK CHI - YUEN ⌋	¶狄志遠議員 ⌋

Figure 7. A sample of misalignment using pure length criteria.

In summary, we have found that the statistical correlation of sentence lengths has a far greater variance for our English-Chinese materials than with the Indo-European materials used by Gale and Church (1991). Despite this, the pure length-based method performs surprisingly well, except for its weakness in handling long stretches of sentences with close lengths.

5. Statistical Incorporation of Lexical Cues

To obtain further improvement in alignment accuracy requires matching the passages' lexical content, rather than using pure length criteria. This is particularly relevant for the type of long mismatched stretches described above.

Previous work on alignment has employed either solely lexical or solely statistical length criteria. In contrast, we wish to incorporate lexical criteria without giving up the statistical approach, which provides a high baseline performance.

Our method replaces equation 5 with the following approximation:

$$\Pr(L_1 \rightleftharpoons L_2 | L_1, L_2) \;\approx\; \Pr(L_1 \rightleftharpoons L_2 | l_1, l_2, v_1, w_1, v_2, w_2, \ldots, v_n, w_n) \tag{9}$$

where

$$v_i = \#\text{occurrences}(\text{English cue}_i, L_1)$$

and

$$w_i = \#\text{occurrences}(\text{Chinese cue}_i, L_2)$$

Again, the dependence is encapsulated within difference parameters δ_i as follows:

$$\begin{aligned}&\Pr(L_1 \rightleftharpoons L_2 | L_1, L_2) \approx \\ &\quad \Pr(L_1 \rightleftharpoons L_2 | \delta_0(l_1, l_2), \delta_1(v_1, w_1), \delta_2(v_2, w_2), \ldots, \delta_n(v_n, w_n))\end{aligned} \tag{10}$$

Bayes' Rule yields

$$\begin{aligned}&\Pr(L_1 \rightleftharpoons L_2 | \delta_0, \delta_1, \delta_2, \ldots, \delta_n) \propto \\ &\quad \Pr(\delta_0, \delta_1, \delta_2, \ldots, \delta_n | L_1 \rightleftharpoons L_2) \Pr(L_1 \rightleftharpoons L_2)\end{aligned} \tag{11}$$

The prior $\Pr(L_1 \rightleftharpoons L_2)$ is evaluated as before. We assume all δ_i values are approximately independent, giving

$$\Pr(\delta_0, \delta_1, \delta_2, \ldots, \delta_n | L_1 \rightleftharpoons L_2) \;\approx\; \prod_{i=0}^{n} \Pr(\delta_i | L_1 \rightleftharpoons L_2) \tag{12}$$

The same dynamic programming optimization can then be used. However, the computation and memory costs grow linearly with the number of lexical cues. This may not seem expensive until one considers that the pure length-based method only uses resources equivalent to that of a single lexical cue. It is in fact important to choose as few lexical cues as possible to achieve the desired accuracy.

Given the need to minimize the number of lexical cues chosen, two factors become important. First, a lexical cue should be highly reliable, so that violations, which waste the additional computation, happen only rarely. Second, the chosen lexical cues should occur frequently, since computing the optimization over many zero counts is not useful. In general, these factors are quite domain-specific, so lexical cues must be chosen for the particular corpus at hand. Note further that when these conditions are met, the exact probability distribution for the lexical δ_i parameters does not have much influence on the preferred alignment.

The bilingual correspondence lexicons we have employed are shown in Figures 3 and 4. These lexical items are quite common in the LegCo domain. Items like "C.B.E." stand for honorific titles such as "Commander of the British Empire"; the other cues are self-explanatory. The cues nearly always appear 1-to-1 and the differences δ_i therefore have a mean of zero. Given the relative unimportance of the exact distributions, all were simply assumed to be normally distributed with a variance of 0.07 instead of sampling each parameter individually. This variance is fairly sharp, but nonetheless, conservatively reflects a lower reliability than most of the cues actually possess.

Table 3. Lexicon employed for paragraph alignment.

:	:
governor	總督

Table 4. Lexicon employed for sentence alignment.

C.B.E.	C.B.E.	C.M.G.	C.M.G.	I.S.O.	I.S.O.
J.B.E.	J.B.E.	J.P.	J.P.	K.B.E.	K.B.E.
L.V.O.	L.V.O.	O.B.E.	O.B.E.	M.B.E.	M.B.E.
Q.C.	Q.C.	January	一月	February	二月
March	三月	April	四月	May	五月
June	六月	July	七月	August	八月
September	九月	October	十月	November	十一月
December	十二月	Monday	星期一	Tuesday	星期二
Wednesday	星期三	Thursday	星期四	Friday	星期五
Saturday	星期六	Sunday	星期日		

Using the lexical cue extensions, the Type I results on the same test file rise to 92.1% of true $L_1 \rightleftharpoons L_2$ pairs correctly identified, as compared to 86.4% for the pure length-based method. The improvement is entirely in the introductory session header. Without the header, the rate is 95.0% as compared to 95.2% earlier (the discrepancy is insignificant and is due to somewhat arbitrary decisions made on anomalous regions). Again, caution should be exercised in interpreting these percentages.

By the alternative Type II measure, 96.1% of the output 1-for-1 pairs were true matches, compared to 86.1% using the pure length-based method. Again, there is an insignificant drop when the header is discarded, in this case from 96.3% down to 95.8%.

Figure 8 shows a sample of the output from alignment with the addition of lexical criteria, for the same data as in Figure 7.

1.	1 HONG KONG LEGISLATIVE COUNCIL – 8 October 1992 ⌋	立法局會議過程正式紀錄 ⌋
2.	¶HONG KONG LEGISLATIVE COUNCIL – 8 October 1992 1 OFFICIAL RECORD OF PROCEEDINGS ⌋	¶一九九二年十月八日星期四 ⌋
3.	¶Thursday, 8 October 1992 The Council met at half - past Two o'clock ⌋	¶下午二時三十分開始會議 ⌋
4.	¶PRESENT ⌋	¶出席者: ⌋
5.	¶THE PRESIDENT HIS EXCELLENCY THE GOVERNOR THE RIGHT HONOURABLE CHRISTOPHER FRANCIS PATTEN ⌋	¶總督彭定康先生(主席) ⌋
6.	¶THE DEPUTY PRESIDENT THE HONOURABLE JOHN JOSEPH SWAINE, C.B.E., Q.C., J.P. ⌋	¶副主席施偉賢議員, C.B.E., Q.C., J.P. ⌋
7.	¶THE CHIEF SECRETARY THE HONOURABLE SIR DAVID ROBERT FORD, K.B.E., L.V.O., J.P. ⌋	¶布政司霍德爵士議員, K.B.E., L.V.O., J.P. ⌋
8.	¶THE FINANCIAL SECRETARY THE HONOURABLE NATHANIEL WILLIAM HAMISH MACLEOD, C.B.E., J.P. ⌋	¶財政司麥高樂議員, C.B.E., J.P. ⌋
9.	(37 correctly aligned matchings omitted)	
10.	¶THE HONOURALBE MAN SAI - CHEONG ⌋	¶文世昌議員 ⌋
11.	¶THE HONOURABLE STEVEN POON KWOK - LIM THE HONOURABLE HENRY TANG YING - YEN, J.P. ⌋	¶潘國濂議員 唐英年議員, J.P. ⌋
12.	¶THE HONOURABLE TIK CHI - YUEN ⌋	¶狄志遠議員 ⌋

Figure 8. A sample of alignment output with lexical criteria incorporated.

Thus, we obtained significant performance improvement by hybridizing lexical and length-based alignment methods within a common statistical framework. Though this is particularly useful for non-alphabetic languages where character length is not as discriminating a feature, we believe improvements will result even when applied to alphabetic languages.

The method was used to produce aligned sentence pairs for the subsequent word translation extraction stage. Only 1-for-1 sentence translations were retained. The remaining data included approximately 17.9Mb of English (about 3 million words) and 9.6Mb of Chinese. The 96% precision on the 1-for-1 sentence pairs turned out to be quite sufficient for the learning of translation lexicons.

6. Word Translation Training Procedure

The Chinese sentences must be segmented before word translation training, because written Chinese consists of a character stream with no space separators between words. Without segmentation, we would not know which Chinese character sequences are legitimate target chunks for translation. Segmentation is a somewhat arbitrary task though, since nearly all individual characters can be considered standalone words; the distinction between Chinese words, compounds, and collocations is unclear and may well be meaningless. In an attempt to circumvent this during earlier phase of the project, we experimented with learning translation associations between English words and individual Chinese characters; while the results were encouraging, they were clearly unsatisfactory.

To segment the Chinese text, therefore, we used an online wordlist in conjunction with an optimization procedure described in (Wu and Fung, 1994). Punctuation is separated out into word-level tokens as a byproduct of this process. According to the word segmentation produced by this method, the Chinese text consists of approximately 3.2 million words or tokens.

In addition, we wished to reduce noise from extraordinarily long sentences, which tend not to be translated sentence-by-sentence. We therefore removed all sentence pairs where either the English sentence was over 70 words long, or the Chinese sentence was over 90 words long. The English text was also normalized for punctuation, raising the word count to about 3.3 million tokens. After all prefiltering, the total input training text for the translation lexicon learning stage consisted of approximately 17.7Mb of English and 12.2Mb of Chinese.

The bilingual training process employs a variant of the model in (Brown et al., 1993), and as such is based on an iterative EM (expectation-maximization) procedure for maximizing the likelihood of generating the Chinese corpus given the English portion. The output of the training process is a set of potential Chinese translations for each English word, together with the probability estimate for each translation. The basic model of Brown *et al.* assumes that the probability of translating a given English sentence $\mathbf{e} = e_1 e_2 \cdots e_l$ into a Chinese sentence $\mathbf{c} = c_1 c_2 \cdots c_m$

following a particular word alignment $\mathbf{a} = a_1 a_2 \cdots a_m$ can be approximated by

$$Pr(\mathbf{c}, \mathbf{a}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(c_j|e_{a_j}) \tag{13}$$

where $t(\cdot)$ are translation probabilities for individual word pairs and ϵ is a small constant. Under this assumption, the expected number of times that any particular word e in an English training sentence $\mathbf{e}$ generates any particular word c in the corresponding Chinese training sentence $\mathbf{c}$ is given by

$$c(c|e; \mathbf{c}, \mathbf{e}) = \frac{t(c|e)}{t(c|e_0) + \cdots + t(c|e_l)} \sum_{j=1}^{m} \delta(c, c_j) \sum_{i=0}^{l} \delta(e, e_i) \tag{14}$$

and the translation probabilities are given by

$$t(c|e) = \lambda_e^{-1} c(c|e; \mathbf{c}, \mathbf{e}) \tag{15}$$

where

$$\lambda_e = \sum_{c} \sum_{\mathbf{c},\mathbf{e} \in \text{corpus}} c(c|e; \mathbf{c}, \mathbf{e}) \tag{16}$$

is the Lagrange multiplier for word e.

The training algorithm treats equations 14—16 as re-estimation formulae for an iterative algorithm as follows:

1. Choose any set of consistent initial values for $t(\cdot)$.
2. Compute the counts for all word translation pairs using equation 14, summing over all sentence pairs in the corpus.
3. Compute the Lagrange multiplier for each English word using equation 16.
4. Re-estimate the translation probabilities using equation 15.
5. Repeat 2—4 until the translation probabilities converge.

For the corpus described here, training time is quite reasonable—approximately 24 hours on a Sparc 10/51— to learn the Chinese translations for a total of 6,536 unique English words prior to filtering. A sample of the output is shown in Figure 5. The most immediate practical use of the output is to have the lexicographer manually delete incorrect entries to produce a translation lexicon. Our first rough evaluation of learning performance is taken with respect to this application: it measures the percentage of English words for which a correct Chinese translation is found within the learned translation set. Of course this measure is meaningful only if the average size of the translation sets is fairly small. We therefore first prune the translation sets with the filters discussed in Section 7, which eliminates many of the low-probability translations and thereby reduces the average size of the translation sets to 2.33 candidates per English word. Even after pruning, the resulting

Table 5. Examples of unfiltered output with probabilities. Note that ⟨⟩ is a special token lumping together all low-frequency Chinese words; *Censorship* is not correctly learned.

I	
0.947	我
0.017	想
0.009	謹此
0.008	希望
0.006	相信
0.003	提出
0.003	認為
0.003	支持
0.002	副主席先生

not	
0.498	不
0.121	不會
0.091	並非
0.090	並
0.072	沒有
0.062	無
0.022	但
0.014	是
0.010	而
0.008	應
0.005	這些
0.005	認為
0.001	他們

other	
0.818	其他
0.117	及其他
0.031	方面
0.011	和
0.011	本港
0.006	與
0.006	國家

threaten	
0.247	威脅
0.119	影響
0.103	構成
0.087	不
0.082	都
0.082	道
0.072	因
0.062	度
0.049	亦
0.027	安定
0.026	嚴
0.023	重
0.016	問題
0.005	父母

UK	
0.234	英國
0.189	⟨⟩
0.114	香港
0.049	處理
0.049	調查
0.049	美國
0.049	and
0.033	大嶼山
0.032	大
0.032	幹線
0.028	b
0.022	監督
0.020	期
0.019	研究
0.018	設計
0.017	水
0.017	北
0.016	情況
0.009	發展
0.005	5

Censorship	
0.314	⟨⟩
0.206	1988年
0.130	錄影帶
0.083	檢查
0.050	明
0.040	管制
0.036	影
0.030	報告
0.032	香港
0.023	一
0.020	根據
0.015	電
0.014	7
0.007	碟

percentage correct is very high—95.1%—as from a randomly drawn sample of 204 English words.[3]

Encouraged by this result, we proceeded toward learning without manual correction. As a first pass, we evaluated the precision of the lexicon obtained by retaining only the *single most probable* translation for each English word. Another randomly drawn sample of 200 words yielded a precision estimate of 91.2% for this. This simple algorithm shows that even fully automatic procedures for English and Chinese are feasible with high precision.

Of course, most words have multiple potential translations, and the above method discards many correct alternative translations of English words. The problem is to find an automatic method for retaining these alternatives, without also retaining an excessive proportion of incorrect translations.

7. Significance Filtering

The training procedure described above results in many translation entries with small or negligible probabilities, which should be pruned to produce a useful lexicon. An obvious solution to reduce noisy lexical entries is to set thresholds on probability. However, absolute thresholds work poorly. Sparse data causes many wrong entries to have inappropriately high probabilities. Conversely, some words are genuinely ambiguous and therefore legitimately spread out the probability across many translations. These cases should not be pruned by absolute thresholds.

We therefore use two filtering criteria that simultaneously penalize for sparse data, and relax for ambiguous words. First, only English words that occur more than 25 times in the corpus are included in the lexicon. Second, for each word, only the translations accounting for the top 0.75 of the probability mass are retained; moreover, any translation with probability less than 0.11 is eliminated. In effect, the filtering threshold rises with data sparseness, and falls with the word's translation entropy.

Evaluating the precision of this approach is slightly more involved, since alternative translation candidates have unequal probabilities. Note that after filtering, the probabilities of the candidates that remain are renormalized to sum to unity for each English word. We use these renormalized probabilities to weight the count of correct translations for the precision estimate. For example, if the translation set for English word *detect* has the two correct Chinese candidates 偵查 with 0.533 probability and 發現 with 0.277 probability, and the incorrect candidate 有關 with 0.190 probability, then we count this as 0.810 correct translations and 0.190 incorrect translations.

Again, another random sample of 200 words was drawn, yielding a weighted precision estimate of 86.0%. Though less than the 91.2% figure for the earlier single-most-probable procedure, this precision is still quite high, even though each English word now has on average 2.33 Chinese translations. Some examples are shown in Figure 6.

Table 6. Examples of filtered output with probabilities.

detect	偵查.533	發現.277	有關.190
Agreement	協議.790	協定.210	
comes	⟨⟩.348	有.334	後.318
brain	人.342	港.335	人才.323
counter	⟨⟩.648	與.352	
trouble	⟨⟩1.000		
empty	只是.615	空.385	
Her	英國.724	政府.276	
beds	病床.687	張.313	

8. Learning Collocation Translations

In the above experiments, the Chinese portion of the corpus was pre-segmented using an existing hand-coded wordlist, i.e., prior linguistic knowledge. Our second attack was to attempt to acquire a translation lexicon completely from scratch. That is, from pure, unsegmented Chinese text, we attempt to learn translations for the English words into appropriate Chinese character sequences of arbitrary form.

One motivation for this "zero-initial-knowledge" experiment is to move closer toward models of language acquisition.

The more practical motivation is to avoid mislearning translations of Chinese compounds as if their constituents were different senses of a polysemous English word. Observe that many of the terms in our corpus are compounds that are not found in the wordlist, such as 通脹 (inflation), with the result that in the previous experiment many English words end up being translated to the multiple constituent words of the Chinese compound. For example, the translation learned for *inflation* is 通 (*tōng*, through) with 0.501 probability, and 脹 (*zhàng*, swell) with 0.499 probability.

To this end, we first induced a Chinese lexicon from the corpus using *CXtract*, a tool we have described elsewhere (Fung and Wu, 1994; Wu and Fung, 1994) and which we extended from *Xtract* (Smadja, 1993). CXtract is an automatic word/collocation extractor for Chinese that employs statistical methods in conjunction with some simple morphological and closed-class syntactic filters. Training on our corpus produced 5,505 multiple-character sequences that were entered into a new, fully automatically-learned Chinese wordlist. A small excerpt is shown in Figure 7. In addition, all individual characters were considered potential words. This lexicon induced by CXtract was then used to segment the Chinese text, producing demarcations of significantly longer segments than in the first experiment. In fact, the total number of segments in the Chinese is reduced by 6.5%, due to recognition of the longer, domain-specific terms. After removing words occurring fewer than 25 times, the corpus contained a total of 5,751 unique Chinese words including single-character forms.

This experiment produced still better results. When the translation lexicon was trained on the newly segmented corpus (again following the procedure of Section 6),

Table 7. Examples of lexical entries produced by CXtract (glosses are hand-produced).

白皮書	(White Paper)
行政局	(Executive Council)
工商界	(Industry and Trade)
保安司	(Security Secretary)
一九九七年	(the year 1997)
立法局選舉	(Election of the Urban Council)
醫院管理局	(Hospital Administration Committee)
警務處處長	(Police Chief)
中英聯合聲明	(Sino-British Joint Declaration)
刑事罪行條例	(Criminal Law Bill)
多議席單票制	(many-seats one-vote system)
機場核心工程	(Airport Core Project)
委員會審議階段	(examining period of the committee)

the manually post-filtered accuracy rose slightly from 95.1% in the first experiment to 96.0%, with an average of 2.25 Chinese translations for each English word. Likewise, whereas the first experiment produced 91.2% correct single-most-probable translations, we obtained a single-most-probable translation precision of 93.5% with the new method.

With significance filtering, the weighted precision rose from 86.0% to 91.0%, which represents a 35.7% reduction in the error rate. Since the average number of Chinese translations per English word fell slightly from 2.33 to 2.25, the new method reduces the number of incorrect "noisy" translations retained. Some examples are shown in Figure 8. There was marked improvement on *brain* and *empty*, where the probabilities for the correct translations had previously been too low. For *comes*, *counter*, and *trouble*, correct translations were learned where the one-stage method had failed to find them at all. *Her* was incorrectly learned by both methods because the capitalized form is used predominantly in phrases such as *Her Majesty's government*, creating high correlations with 英國 (England) and 政府 (government). A larger excerpt of the 200 evaluated translation samples can be found in the appendix.

Table 8. Examples of filtered output with probabilities, from the two-stage procedure.

comes	來.511	到.489	
brain	人才1.000		
counter	反.498	對.276	付.226
trouble	煩.415	麻.313	事.272
empty	空1.000		
Her	英國.747	政府.253	
beds	病床.573	張.262	床.165
do	我1.000		
an	,.417	的.397	一個.186
immigration	入境.812	出.188	
final	最後.732	最終.268	

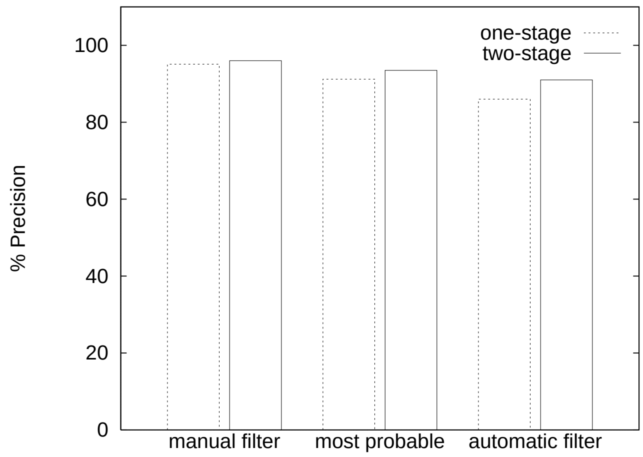

Figure 9. Summary of precision results, showing consistent improvement obtained with automatically-derived monolingual lexicon.

Much of the improvement in performance appears to be due to the fact that correct identification of collocations in the Chinese text facilitates closer matchings to the English words. For example, the two-stage method correctly learned the top translation candidate for *inflation* as 通脹 with 0.809 probability, instead of the separate translations 通 and 脹 seen in the one-stage method. Moreover, it assigned 0.191 probability to 通貨膨脹 as well, a collocation made up of 通貨 (*tōnghuò*, currency circulation) and 膨脹 (*péngzhàng*, swell). This is actually the full unabbreviated form, from which the more convenient 通脹 is derived. Learning translation patterns for word sequences is also a goal of the later models of Brown *et al.* (1993), but because those models attempt to learn both the collocations and translations simultaneously, the computational cost is quite heavy. In contrast, the two-stage learning procedure used here first employs bottom-up monolingual training, followed by bilingual training, achieving excellent performance with relatively low cost. A summary graph comparing the precision results is shown in Figure 9.

9. Conclusion

The series of experiments reported here are, to our knowledge, the first large-scale empirical demonstrations of the applicability of pure statistical techniques to bilingual English-Chinese lexical acquisition, and possibly the first between Indo-European and non-Indo-European languages. We have obtained high precision rates, between 86.0% and 96.0%, for lexicons of significant size, around 6,500 English words, using significance filtering. The learned lexicons are well-adapted to the corpus domain. Accuracy on paragraph and sentence alignment is improved significantly by incorporating a restricted set of manually-derived lexical cues into the probabilistic optimization, achieving accuracies in the 96% range. For word translation, a two-stage monolingual-bilingual statistical learning procedure outperforms a hybrid procedure using a pre-existing machine-readable lexicon; it yields excellent performance on translation acquisition for words, compounds, and collocations, at reasonable computational cost.

Acknowledgments

We are indebted to Bill Gale, Ken Church, and Pascale Fung for helpful clarifying discussions. Eva Fong, Cindy Ng, and Linda Peto have also contributed significantly to the general development of the SILC project. The online Chinese wordlist (BDC, 1992) was provided by Behavior Design Corporation.

Appendix

The following is a randomly drawn subset of the 200 samples used evaluate the two-stage learned translation lexicon.

will	會.413 ,.223 將.193 將會.169
government	政府1.000
person	人.567 人士.432
issues	問題.762 事項.237
low	低1.000
thus	因此.497 因而.274 使.228
port	港口1.000
lot	很多.588 不少.411
Could	?.438 可否.346 請問.215
region	區.350 亞.239 地區.236 區內.173
stress	強調1.000
assets	資産1.000
history	歷史.637 上.199 史.163
completely	完全1.000
assurance	保證1.000
More	更.674 更多.325
nevertheless	不過.434 但.365 雖然.200
controversial	爭議.428 性.217 爭論.180 具.173
stressed	強調1.000
freedoms	自由1.000
unattended	獨.477 留在.274 中.248
welcomed	歡迎.654 受.185 受到.160
establishments	場所.534 機構.465
insist	堅持.730 要.269
deserves	值得1.000
select	選.480 專責委員會.296 挑.222
gradual	逐步.215 循.170 漸.159 進.154 序.152 逐漸.148
nowadays	現時.534 今.465
remarkable	出.596 傑.403
Recent	最近.614 近期.385
chief	總.629 裁.370
aiming	目標.316 目.284 就是.209 各項.189
workshops	工場1.000
temporarily	暫時.581 臨時.418
ongoing	不斷.412 進行.363 持續.223
Trustee	人.387 受.306 託.306
loud	聲.576 大.423
write	寫1.000
cents	仙.652 每.347
headed	首.374 領導.312 冷靜.312

Notes

1. Cantonese is one of the four Chinese languages. Written Cantonese employs the same characters as Mandarin, with some additions. Though there are grammatical and usage differences between the Chinese languages, as between German and Swiss German, the written forms can be read by all Chinese.
2. The difference is also partly due to the fact that Gale and Church plot paragraph lengths instead of sentence lengths. We have chosen to plot sentence lengths because that is what the algorithm is based on.
3. Person names were excluded, but all other proper names were retained for this evaluation.

References

BDC. 1992. *The BDC Chinese-English Electronic Dictionary (version 2.0)*. Behavior Design Corporation.

Brown, P.F., J. Cocke, S.A. DellaPietra, V.J. DellaPietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):29–85.

Brown, P.F., S.A. DellaPietra, V.J. DellaPietra, and R.L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Brown, P.F., Jennifer C. Lai, and R.L. Mercer. 1991. Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics*, pages 169–176, Berkeley.

Catizone, R., G. Russell, and S. Warwick. 1989. Deriving Translation Data from Bilingual Texts. In *Proceedings of the First International Acquisition Workshop*, Detroit.

Chen, Stanley F. 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, pages 9–16, Columbus, OH.

Church, K.W. 1993. Char-align: A Program for Aligning Parallel Texts at the Character Level. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, pages 1–8, Columbus, OH.

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*, pages 1–8, Columbus, OH, June.

Fung, Pascale and Dekai Wu. 1994. Statistical Augmentation of a Chinese Machine-Readable Dictionary. In *Proceedings of the Second Annual Workshop on Very Large Corpora*, pages 69–85, Kyoto, August.

Gale, W.A. and K.W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics*, pages 177–184, Berkeley.

Gale, W.A., K.W. Church, and D. Yarowsky. 1993. A Method for Disambiguating Word Senses in a Large Corpus. In *Computers and the Humanities*.

Kay, M. and M. Röscheisen. 1988. Text-Translation Alignment. Technical Report P90-00143, Xerox Palo Alto Research Center.

Smadja, F.A. 1993. Retrieving Collocations From Text: Xtract. *Computational Linguistics*, 19(1):143–177.

Smadja, F.A. and K.R. McKeown. 1994. Translating Collocations for Use in Bilingual Lexicons. In *Proceedings of the ARPA Human Language Technology Workshop*, Princeton, N.J., March.

Sperberg-McQueen, C.M. and L. Burnard. 1992. Guidelines for Electronic Text Encoding and Interchange. Version 2 draft.

Wu, Dekai and Pascale Fung. 1994. Improving Chinese Tokenization with Linguistic Filters on Statistical Lexical Acquisition. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 180–181, Stuttgart, October.