In *AMTA-94, Association for Machine Translation in the Americas*, Columbia, MD: Oct. 94, 206–213.

LEARNING AN ENGLISH-CHINESE LEXICON FROM A PARALLEL CORPUS

Dekai Wu
Xuanyin Xia
HKUST
Department of Computer Science
University of Science & Technology
Clear Water Bay, Hong Kong
{dekai,samxia}@cs.ust.hk

Abstract

We report experiments on automatic learning of an English-Chinese translation lexicon, through statistical training on a large parallel corpus. The learned vocabulary size is non-trivial at 6,517 English words averaging 2.33 Chinese translations per entry, with a manually-filtered precision of 95.1% and a single-most-probable precision of 91.2%. We then introduce a *significance filtering* method that is fully automatic, yet still yields a weighted precision of 86.0%. Learning of translations is adaptive to the domain. To our knowledge, these are the first empirical results of the kind between an Indo-European and non-Indo-European language for any significant corpus size with a non-toy vocabulary.

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, from the SILC project at HKUST which is studying machine learning of natural language translation. Specifically, we report accuracy rates for a new performance measure: the precision of statistical translation lexicon acquisition, between English and Chinese.

In the first phase of SILC (*s*tatistical *i*nter-*l*ingual *c*onversion), we have (1) collected a bilingual corpus of parallel English and Chinese text, (2) aligned the sentences within the corpus, and (3) learned a bilingual lexicon from the aligned data, to be embedded within learned translation models. Preliminary results on the first two steps were reported in Wu (1994); an updated summary is given in section 2, before we discuss the bilingual lexicon learning.

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.

2 A Sentence-Aligned Corpus

Though large parallel bilingual corpora are relatively scarce compared with monolingual corpora, they have generated more interesting results. Significant progress has been made on problems including automatic sentence alignment (Kay & Röscheisen 1988; Catizone *et al.* 1989; Gale & Church 1991; Brown *et al.* 1991; Chen 1993), coarse alignment (Church 1993; Fung & Church 1994), statistical machine translation (Brown *et al.* 1990; Brown *et al.* 1993), word alignment (Dagan *et al.* 1993), word sense disambiguation (Gale *et al.* 1993), and collocation learning (Smadja & McKeown 1994), all exploiting parallel 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.

Currently, the material in the corpus consists primarily of fairly tight, literal sentence translations from the parliamentary proceedings of the Hong Kong Legislative Council. The original materials were not designed to be available in machine-readable form, and their obscure format necessitated heavy conversion and reformatting, using both manual and automatic processing. For the experiments reported in this paper, we began with 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).

The corpus is aligned using a hybrid statistical and lexical strategy. A dynamic programming algorithm optimizes an approximation to

$$\arg\max_{\mathcal{A}} \Pr(\mathcal{A}|\mathcal{T}_1, \mathcal{T}_2) \approx \arg\max_{\mathcal{A}} \prod_{(L_1 \rightleftharpoons L_2)\in\mathcal{A}} \Pr(L_1 \rightleftharpoons L_2 | L_1, L_2) \tag{1}$$

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. The approximation depends largely on the lengths of the sentences, with some assistance from a very small set of high-reliability lexical cues:

$$\Pr(L_1 \rightleftharpoons L_2|L_1, L_2) \approx \Pr(L_1 \rightleftharpoons L_2|\delta_0(l_1, l_2), \delta_1(v_1, w_1), \ldots, \delta_n(v_n, w_n)) \tag{2}$$

where l_1 and l_2 are the lengths of the English and Chinese passages; v_i is the number of occurrences of an English cue in L_1 and w_i is the number of occurrences of the corresponding Chinese cue in L_2; and $\delta_i(\cdot)$ is a normalized difference function that encapsulates the dependence within a single parameter for each i. For details, the reader is referred to Wu (1994). The algorithm produces alignments with about 92% recall accuracy; an example is shown in Figure 1.

Some of the sentence alignments match different numbers of English and Chinese sentences, as in line 3 of Figure 1. Such cases were discarded because they tend to be looser translations, and are thus harder to extract accurate correlations from. After retaining only 1-for-1 sentence translations, the remaining data included approximately 17.9Mb of English (about 3 million words) and 9.6Mb of Chinese. The precision on the 1-for-1 sentence pairs was approximately 96%, which turned out to be quite sufficient for the subsequent learning of translation lexicons.

3 Training Procedure

The Chinese portion of the corpus must be segmented before 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 an earlier phase

1. I would like to talk about public assistance. ⌋	我想談及公共援助問題。⌋
2. 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%。⌋
3. 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。⌋
4. 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。⌋

Figure 1: A sample of length-based alignment output.

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 (BDC 1992) in conjunction with an optimization procedure described in Wu & 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 Brown *et al.*'s (1993) model, 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.* (1993) 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

$$Pr(\mathbf{c}, \mathbf{a} | \mathbf{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(c_j | e_{a_j}) \tag{3}$$

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

$$c(c|e; \mathbf{c}, \mathbf{e}) = \frac{t(c|e)}{t(c|e_0) + \cdots + t(c|e_l)} \sum_{j=1}^{m} \delta(c, c_j) \sum_{i=0}^{l} \delta(e, e_i) \tag{4}$$

and the translation probabilities are given by

$$t(c|e) = \lambda_e^{-1} c(c|e; \mathbf{c}, \mathbf{e}) \tag{5}$$

where

$$\lambda_e = \sum_c \sum_{\mathbf{c},\mathbf{e} \in \text{corpus}} c(c|e; \mathbf{c}, \mathbf{e}) \tag{6}$$

is the Lagrange multiplier for word e.

The training algorithm treats Equations 4—6 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 4, summing over all sentence pairs in the corpus.
3. Compute the Lagrange multiplier for each English word using Equation 6.
4. Re-estimate the translation probabilities using Equation 5.
5. Repeat 2—4 until the translation probabilities converge.

4 Baseline Performance

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 2.

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 5, 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 percentage correct is very high—95.1%—as estimated from a randomly drawn sample of 204 English words.[1]

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.

[1] Person names were excluded, but all other proper names were retained for this evaluation.

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

Figure 2: Examples of unfiltered output with probabilities. Note that ⟨⟩ is a special token lumping together all low-frequency Chinese words; *Censorship* is not correctly learned.

5 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 introduce two *significance 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 typical examples are shown in Figure 3, where the highly-ranked translations of *detect* and *Agreement* are correct. Figure 3 also shows a range of the types of errors made. For words that occur frequently within frozen collocations, such as *brain* (as in *brain drain*) or *empty*, the probabilities for the correct translations can be too low. *Her* was incorrectly learned 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). Alternatively, the entire collocations may be learned, as in the case of *beds* being translated to 病床 (*hospital bed*, *sickbed*). With less frequent words or inflectional forms such as *counter*, *comes*, and *trouble*, the correct translations

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

Figure 3: Examples of significance-filtered output with probabilities.

sometimes fail to be learned at all. Yet the large majority of entries followed the pattern exemplified by *detect* and *Agreement*.

Adaptiveness is one key strength of the statistical acquisition method; i.e., the translation learning is sensitive to the domain of translation. Translations for many governmental terms were picked up that would not normally be found in a hand-constructed translation lexicon. A more subtle example is *brain*, for which a correct translation 人才 was learned although the ordinary translation 腦 was not. In the domain of political discourse, the usage of *brain* is less often in a biological sense than in *brain drain*, for which 人才外流 is the appropriate translation.

6 Conclusion

We have described a series of techniques for automatically extracting a bilingual English-Chinese translation lexicon. The experiments reported here are, to our knowledge, the first large-scale empirical demonstrations of the applicability of pure statistical techniques to this task, and possibly the first between Indo-European and non-Indo-European languages. We have obtained high precision rates, between 86.0% and 95.1%, for lexicons of non-toy size, around 6,500 English words. Moreover, we have introduced a *significance filtering* method that can feed translation hypotheses of high precision into a manual post-filtering stage, or alternatively, yields excellent results even in fully automatic mode. A graph summarizing the precision results is shown in Figure 4.

The statistical methods described are able to improve translation accuracy because the learned lexicons are finely adapted to the corpus domain. However, the methods still possess a strong non-adaptive component in the large hand-coded Chinese wordlist. Space limitations preclude description of an additional series of experiments that demonstrate the potential of adaptive techniques even more strongly. In the full paper (Wu & Xia 1994), we show that even when the hand-coded wordlist is discarded, a fully adaptive procedure still raises performance significantly.

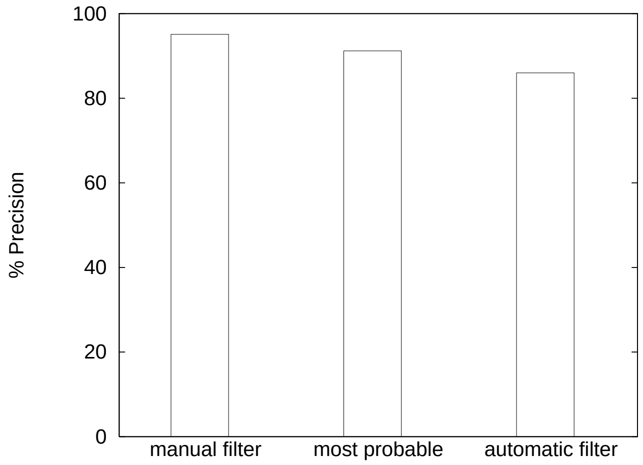

Figure 4: Summary of precision results.

Acknowledgements

We are indebted to Kenneth Church, William Gale, Pascale Fung, and the anonymous reviewers for helpful comments and suggestions, and to Eva Fong, Cindy Ng, and Derek Ngok for general contributions in the SILC project. The online Chinese wordlist (BDC 1992) was provided by Behavior Design Corporation.

References

BDC. 1992. *The BDC Chinese-English electronic dictionary (version 2.0)*. Behavior Design Corporation.

BROWN, PETER F., JOHN COCKE, STEPHEN A. DELLAPIETRA, VINCENT J. DELLAPIETRA, FREDERICK JELINEK, JOHN D. LAFFERTY, ROBERT L. MERCER, & PAUL S. ROOSSIN. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):29–85.

BROWN, PETER F., STEPHEN A. DELLAPIETRA, VINCENT J. DELLAPIETRA, & ROBERT L. MERCER. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

BROWN, PETER F., JENNIFER C. LAI, & ROBERT L. MERCER. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics*, 169–176, Berkeley.

CATIZONE, ROBERTA, GRAHAM RUSSELL, & SUSAN 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*, 9–16, Columbus, OH.

CHURCH, KENNETH 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*, 1–8, Columbus, OH.

DAGAN, IDO, KENNETH W. CHURCH, & WILLIAM A. GALE. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora*, 1–8, Columbus, OH.

FUNG, PASCALE & KENNETH W. CHURCH. 1994. K-vec: A new approach for aligning parallel texts. In *Proceedings of the Fifteenth International Conference on Computational Linguistics*, 1096–1102, Kyoto.

FUNG, PASCALE & DEKAI WU. 1994. Statistical augmentation of a Chinese machine-readable dictionary. In *Proceedings of the Second Annual Workshop on Very Large Corpora*, 69–85, Kyoto.

GALE, WILLIAM A. & KENNETH W. CHURCH. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics*, 177–184, Berkeley.

GALE, WILLIAM A., KENNETH W. CHURCH, & DAVID YAROWSKY. 1993. A method for disambiguating word senses in a large corpus. In *Computers and the Humanities*.

KAY, MARTIN & M. RÖSCHEISEN. 1988. Text-translation alignment. Technical Report P90-00143, Xerox Palo Alto Research Center.

SMADJA, FRANK A. & KATHLEEN R. MCKEOWN. 1994. Translating collocations for use in bilingual lexicons. In *Proceedings of the ARPA Human Language Technology Workshop*, Princeton, N.J.

WU, DEKAI. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics*, 80–87, Las Cruces, New Mexico.

WU, DEKAI & PASCALE FUNG. 1994. Improving Chinese tokenization with linguistic filters on statistical lexical acquisition. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, Stuttgart. To appear.

WU, DEKAI & XUANYIN XIA, 1994. Large-scale automatic extraction of an English-Chinese lexicon. Submitted.