A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora

Pascale Fung

Human Language Technology Center
Department of Electrical and Electronic Engineering
University of Science and Technology (HKUST)
Clear Water Bay, Hong Kong
pascale@ee.ust.hk

Abstract. We present two problems for statistically extracting bilingual lexicon: (1) How can noisy parallel corpora be used? (2) How can non-parallel yet comparable corpora be used? We describe our own work and contribution in relaxing the constraint of using only clean parallel corpora. **DKvec** is a method for extracting bilingual lexicons, from noisy parallel corpora based on arrival distances of words in noisy parallel corpora. Using **DKvec** on noisy parallel corpora in English/Japanese and English/Chinese, our evaluations show a 55.35% precision from a small corpus and 89.93% precision from a larger corpus. Our major contribution is in the extraction of bilingual lexicon from non-parallel corpora. We present a first such result in this area, from a new method–**Convec**. **Convec** is based on context information of a word to be translated. We show a 30% to 76% precision when top-one to top-20 translation candidates are considered. Most of the top-20 candidates are either collocations or words related to the correct translation. Since non-parallel corpora contain a lot more polysemous words, many-to-many translations, and different lexical items in the two languages, we conclude that the output from **Convec** is reasonable and useful.

1 Introduction

Bilingual lexicon extraction using large corpora is a relatively new area of research where the curve of progress has been positively steep since the beginning. The initial breakthrough in the area came from using statistical information of word features from clean, parallel corpora for automatic term translation. A parallel corpus is a pair of translated texts. A clean corpus is one which contains minimal translation noise such as when a sentence in one text is not translated in the other, or when the sentence boundary is not clear. Most parallel corpora are cleaned up by manual preprocessing to eliminate such noise. The most common word feature used is the co-occurrence of words in parallel sentences. Given a bilingual corpus where pairs of translated sentences are aligned, co-occurring words across the two languages in the text are extracted and matched by using correlation measures.

Efforts in using clean, parallel corpora for translation have been met with some objections in the research community. One main objection came from the fact that restricting the resources to clean, parallel corpora is too constraining. Therefore, we want to extract statistical word information from different types of real world data, including parallel but noisy corpora, as well as non-parallel texts of monolingual data from the same domain.

Another constraint of previous algorithms is their implicit reliance on common character sets between parallel corpora. So, another goal for us is to develop robust algorithms which can process languages that do not share etymological roots, such as Chinese and English, or Japanese and English.

In this paper, we present our view and our solutions to these problems. We show how word occurrence frequency, position, distance, context of the words, and dictionary seed words are used in different algorithms targeted at extracting bilingual lexicon from clean parallel corpora, noisy parallel corpora, and from non-parallel yet comparable corpora.

2 Bilingual lexicon extraction from parallel corpora

Algorithms for bilingual lexicon extraction from parallel corpora exploit the following characteristics of translated, bilingual texts:

1. Words have **one sense** per corpus
2. Words have **single translation** per corpus
3. **No missing translations** in the target document
4. **Frequencies** of bilingual word occurrences are **comparable**
5. **Positions** of bilingual word occurrences are **comparable**

Most translated texts are domain-specific, thus their content words are usually used in one sense and are translated consistently into the same target words. Pairs of sentences from both sides of the translated documents contain the same content words, and each word occurs in approximately the same sentences on both sides. Once the corpus is aligned sentence by sentence, it is possible to learn the mapping between the bilingual words in these sentences.

Sometimes lexicon extraction is a by-product of alignment algorithms aimed at constructing a statistical translation model [2–4, 12, 23, 32]. Others such as [6, 7] use an EM-based model to align words in sentence pairs in order to obtain a technical lexicon. Some other algorithms use sentence-aligned parallel texts to further compile a bilingual lexicon of technical words or terms using similarity measures on bilingual lexical pairs [21, 25, 29]. Yet others focus on translating phrases or terms which consist of multiple words [6, 25, 29].

The main inspiration for our work [10, 14], to be described in the following section, comes from [21] who propose using word occurrences patterns and average mutual information and t-scores to find word correspondences as an alternative to the IBM word alignment model. Given any pair of bilingual words, their occurrence patterns in all sentences are transformed into binary occurrence vectors, where the presence of a word in sentence i assigns a 1 to the i-th dimension of the binary vector w.

The correlation between a word pair is then:

$$\begin{aligned} W(w_s, w_t) &= \log_2 \frac{\Pr(w_s = 1, w_t = 1)}{\Pr(w_s = 1)\Pr(w_t = 1)} \\ &= \log_2 \frac{a \cdot (a+b+c+d)}{(a+b)\cdot(a+c)} \end{aligned}$$

A word pair is considered only if their $t > 1.65$ where

$$t \approx \frac{\Pr(w_s = 1, w_t = 1) - \Pr(w_s = 1)\Pr(w_t = 1)}{\sqrt{\frac{1}{a+b+c+d}\Pr(w_s = 1, w_t = 1)}}$$

This work laid the basis for other translation algorithms using correlation scores.

2.1 DKvec: From clean parallel corpora to noisy parallel corpora

Our work in bilingual lexicon extraction is motivated by the need to handle parallel corpora which do not have clear sentence boundaries, which contain many insertion or deletion *noise*, which consist of language pairs across families. These corpora are difficult to align. **DKvec** [10, 13, 14] is an algorithm which, instead of looking at the position vector of a word, looks at the arrival distance vector of the word and compares it with that of target words. An arrival distance vector is a vector whose values are the positional differences between two successive occurrences of the word in the text. It is based on the notion that while similar words do not occur at the exact same position in each half of the corpus, distances between instances of the same word are similar across languages. Dynamic time warping (DTW) is used to find a mapping between arrival distance vectors. Matched word pairs are then used to align the noisy parallel corpus by segments, taking into account insertion and deletions. From the aligned corpus, words are represented in the word position binary vector form and are matched using Mutual Information score, much like the approach in [12, 21]. Fig. 1 shows recency vectors of some words and Fig. 2 shows the DTW path between the words *Governor* in English and Chinese.

We tested the algorithm on a noisy parallel corpus of English and Chinese texts, and another one of English and Japanese texts. Our algorithm yields a 55.35% precision in technical word translation for the small English/Japanese corpus. In the English/Chinese version we collected translations of 626 English words and suggested candidate translations for 95 English term translations. Word translations were evaluated as 89.9% accurate. Our algorithm produces a suggested word list for each technical term in the English/Chinese corpus and improved human translator performance by an average of 47%.

Fig. 1. Recency vector signals showing similarity between *Governor* in English and Chinese, contrasting with *Bill* and *President* in English

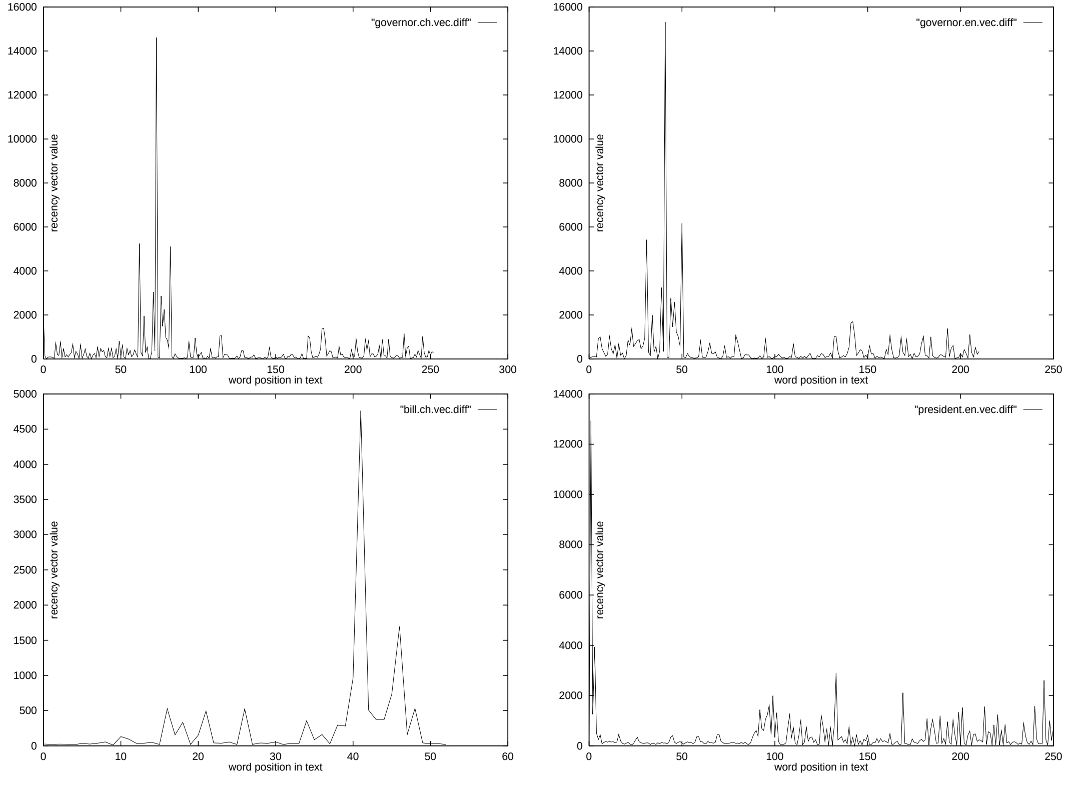
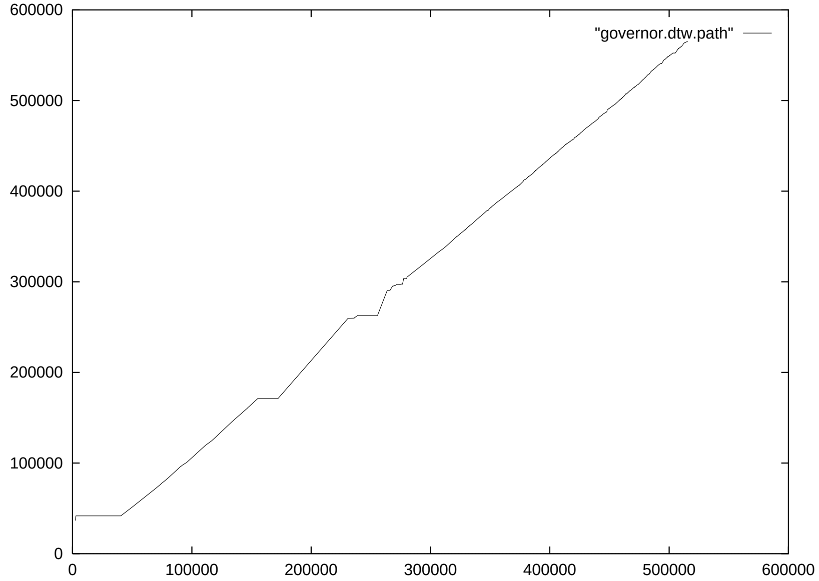


Fig. 2. Dynamic Time Warping path for *Governor* in English and Chinese. The axes represent the lengths of the two texts.



3 Bilingual lexicon extraction from non-parallel corpora

The very trend of using bilingual parallel corpora for machine translation was started by Jelinek's group formerly at IBM. Their work, and others which followed, are based on the conjecture that there must always exist a parallel corpus between any pair of languages for which *"mutual translation is important enough"*. In fact, organizations such as the Linguistic Data Consortium have been dedicating resources to *collect* such parallel corpora in different language pairs. However, acquiring and processing of parallel corpora is labour-intensive and time-consuming. It is still unlikely that one can find parallel corpora in any given domain in electronic form. Whereas even labour-intensive effort is certainly necessary and worthwhile for building baseline bilingual lexicons, we believe that there is an even larger corpora resource which is under-exploited—monolingual texts forming non-parallel yet comparable corpora.

With the advent of Internet technology and the World Wide Web, it has become obvious that such type of non-parallel, comparable corpora are more abundant, more up-to-date, more accessible than parallel corpora. Such corpora can be collected easily from downloading electronic copies of newspapers, journals, articles, and even books from the World Wide Web. At any given time, there are a lot more comparable corpora available than parallel corpora. At any given time, there are always a larger number of new words in comparable corpora than in parallel corpora because there is a delay in "time-to-market" for parallel corpora due to the human translation process. There is a great potential for one to obtain a mapping between bilingual word pairs from these non-parallel but comparable texts in the source language and the target language.

3.1 Comparable corpora

Unlike parallel corpora, which are clearly defined as translated texts, there is a wide variation of *nonparallelness* in monolingual data. Nonparallelness is manifested in terms of differences in *author, domain, topics, time period, language.*

The most common text corpora have nonparallelness in all these dimensions. The higher the degree of nonparallelness, the more challenging is the extraction of bilingual information. Parallel corpora represent the extreme example where all dimensions of nonparallelness except the language are reduced to zero. At the other extreme, newspapers from different time periods have different authors, sometimes cover different domains, and even have very different perspectives on the same events leading to topical differences. Such a corpus is nevertheless a desirable source of bilingual information, especially for new words.

As the nonparallelness of the texts increases, it is more difficult to find statistical usage patterns in the terms. Nonparallelness leads to the following characteristics:

1. Words have **multiple senses** per corpus
2. Words have **multiple translations** per corpus
3. Translations might **not exist** in the target document
4. **Frequencies** of occurrence **not comparable**
5. **Positions** of occurrence **not comparable**

Because of the above, bilingual lexicon extraction from non-parallel corpora is a far more difficult task than that from parallel corpora. Hence, we need a departure from the general paradigm of statistical methods applied to parallel corpora. We have been studying this problem since 1995 [9, 11, 15, 16] and have discovered the following characteristics of comparable corpora:

1. For the same topic, words have **comparable contexts** across languages
2. Words in the same domain and the same time period have **comparable usage** patterns (e.g. Zipf's law)

In the following sections, we describe a new method, **Convec**, which finds translation *candidates* for new words from online newspaper materials. When the source word is unambiguous, the top candidate is usually the correct translation. When the source word is polysemous, or has multiple translations, then the top candidates include other words which are collocations of the correct translation.

3.2 Convec: Moving on to non-parallel corpora

Content information of a monolingual text and context information of its words have been used in algorithms for author characterization from documents [26], document categorization from queries [1, 5, 24, 27, 30] and sense disambiguation between multiple usages of the same word [8, 20, 19, 17, 18, 22, 28, 33]. We propose to use context information of a word to find its counter part in the other language in a comparable corpus.

Our goal is to find translation or translation candidates for new words which are not found in an online dictionary, and then use the result to augment the dictionary, in order to improve the quality of a broad-coverage English-Chinese MT system at the Human Language Technology Center [31].

As an example, while using this MT system, we discovered that it is unable to translate the word *flu* which nevertheless occurs very frequently in the texts. Other words it cannot translate include names of politicians and public figures which only recently start to appear frequently in the news, medical and technical terms which are used in the context of recent events. In the following sections, we use *flu* and the discovery of its translation as a tutorial example to describe our algorithm.

3.3 The algorithm

Our algorithm finds bilingual pairs of words from non-parallel, comparable texts by using an Information Retrieval (IR) approach. It extracts the context of an unknown word in the source language and treats that as a *query*. It then looks at the *documents*—contexts of all candidate translations in the target language, and finds the translation in the *document* which best matches the *query*. Using the example of *flu* → 流感/*liougan*, our algorithm discovers this translation by matching the context of *flu* to that of 流感/*liougan* found in different English and Chinese newspaper articles in Hong Kong. Table 1. shows some occurrences of *flu* in the English newspaper.

Table 1. Contexts of *flu* from English newspaper articles

effect businesses avian	flu	
responsible called bird	flu	resulted
vaccine combat bird	flu	ready
The government handled bird	flu	crisis health
bird	flu	crisis
The deadly bird	flu	spread
VACCINE combat bird	flu	ready summer
THE government handled bird	flu	crisis health
bird	flu	crisis
THE deadly bird	flu	spread
possibility bird	flu	transmitted humans
This bird	flu	able
He cited bird	flu	evidence need change No
After avian	flu	subsides scientists expect
If bird	flu	struck Hong Kong
bird	flu	monetary turbulence soon Moreover
THE bird	flu	caused concern people Hong Kong

Algorithm

1. Precompile context vector models of all unknown words W_e in the source language
2. Precompile context vector models of all candidate translation words W_c in the target language
3. For all W_e, W_c, compute $similarity(W_e, W_c)$.
4. Rank the output according to this similarity score.
5. (I)Choose the N highest ranking W_c as translation candidate for W_e.
6. (II) Choose the M highest ranking (W_e, W_c) as new lexicon entries for the bilingual dictionary

3.4 Finding the TF of words

If we collect all the words in the context of *flu* and count their occurrence frequency **in the context of** *flu*, then we will get the list partially shown in Table 2. This frequency is the word **term frequency (TF)**. The right side of the table shows the most frequent words in the context of 流感/*liougan*. Note the similarity in the two lists.

Table 2. Words in the context of *flu*/流感 are similar.

English	TF	Chinese	TF
bird	284	事件 (event)	218
virus	49	病毒 (virus)	217
people	45	政府 (establishment)	207
Sydney	38	感染 (contraction)	153
scare	32	表示 (denote)	153
spread	19	沒有 (doesn't_exist)	134
deadly	19	病人 (invalid)	106
government	16	專家 (consultancy)	100
China	14	部門 (branch)	96
new	13	染上 (catch)	93
crisis	13	醫院 (hospital)	92
outbreak	12	情況 (circumstance)	90
hospital	12	處理 (deal_with)	89
chickens	9	醫生 (doctor)	49
spreading	8	染上 (infected)	47
prevent	8	醫院 (hospital)	44
crisis	8	沒有 (no)	42
health	8	政府 (government)	41

3.5 Mapping context words using bilingual dictionary

To establish the fact the contexts of *flu/liougan* are similar, we need to find a mapping between the *context words* in the left and right columns of Table 3. This is achieved via an online dictionary—the same one used by the MT system. It goes without saying that only known words are considered in the context vector. So for example, *virus* is found in the dictionary and it is considered a common word between the vectors of *flu/liougan*, whereas even though the name of the Chief Executive appears in both contexts, they are not used as seed words since they are unknown to the dictionary. In total, there are 233 common seed words shared by *flu/liougan*.

Table 3. Some of the 233 common seed words shared by 流感 and *flu*.

Common Index	Chinese word	English word
10614	以往	formerly
10677	出售	for_sale
10773	自由	free
1107	合併	amalgamate
11156	恢復	get_over
11682	香港	H.K.
11934	主管	head
1197	分析	analysis
12229	假日	holiday
12345	醫院	hospital
12635	抵抗力	immunity
12723	好轉	improve
1276	發言人	announcer
12968	之外	infra
13039	調查	inquire_into
12345	醫院	hospital
13845	實驗室	lab
23306	病毒	virus

Table 5. shows that the contexts of *flu* and another unrelated word, *shop* are not similar. Consequently, we found only 41 common seed words shared between *shop/liougan*.

3.6 Finding the IDF of words

We could proceed to compute the similarity of context vectors based on the frequency of the common words they share. However, there is a problem with using TF only. Note that in the example of *flu/liougan*, their common words, such as *virus, infection, spread* are mostly content words highly related to the usage of *flu/liougan* themselves, whereas words such as *discuss, fall, forward, frequently* which are shared by the context vectors of *shop/liougan* have much more general usage. We want to emphasize the significance of common words such as *virus, infection, spread* and deemphasize that of *discuss, fall, forward* by using another frequency, the Inverse Document Frequency (IDF). IDF accounts for the overall occurrence frequency of a context words in the *entire corpus*:

Table 4. Words in the context of *shop/*流感 are not similar.

English	TF	Chinese		TF
pet	5	事件	(event)	218
nine	4	病毒	(virus)	217
here	4	政府	(establishment)	207
ago	4	感染	(contraction)	153
stopping	3	表示	(denote)	153
six	3	沒有	(doesn't_exist)	134
reduce	3	病人	(invalid)	106
owner	3	專家	(consultancy)	100
business	3	部門	(branch)	96
walk	2	染上	(catch)	93
terminate	2	醫院	(hospital)	92
talk	2	內地	(inland)	91
square	2	情況	(circumstance)	90
space	2	處理	(deal_with)	89

Table 5. Some of the 41 common seed words shared by 流感 and *shop*.

Common Index	Chinese word	English word
10363	討論	discuss
13048	下降	fall
14472	引言	forward
14756	不時	frequently
15357	才	gift
15582	過去的	gone
17960	有關連的	incident
17995	包	include
18475	內	inside
20818	本地	local
21699	出售	market
22737	月	month
25320	物主	owner
26245	寵物	pet
26987	小灘	pool
27032	一般的	popular
27413	提出控告	prefer
27633	五月	price

$$\begin{aligned} \text{IDF} &= \log \frac{n_{max}}{n_i} + 1 \\ \text{where } \; n_{max} &= \text{the maximum frequency of} \\ & \text{any word in the corpus} \\ n_i &= \text{the total number of occurrences} \\ & \text{of word } i \text{ in the corpus} \end{aligned}$$

3.7 Similarity measures

Now we can visualize the context vector for *flu* to have the dimension of the bilingual dictionary we use, in this case, 20,000. The i-th dimension of this vector is $w_i = TF_i \times IDF_i$. It is zero if the i-th word does not appear in the context of *flu*. Similarly we obtain the context vectors of all unknown words in the source language and context vectors of all candidate words in the target language.

To locate translation candidates for *flu*, we have to compare the context vector of *flu* with context vectors of all Chinese words. The most common similarity measure used in the IR community is the Cosine Measure. We use one of the variants of the Cosine Measure:

$$\begin{aligned} S(W_c, W_e) &= \frac{\Sigma_{i=1}^{t}(w_{ic} \times w_{ie})}{\sqrt{\Sigma_{i=1}^{t} {w_{ic}}^2 \times \Sigma_{i=1}^{t} {w_{ie}}^2}} \\ \text{where } \; w_{ic} &= TF_{ic} \times IDF_i \\ w_{ie} &= TF_{ie} \times IDF_i \end{aligned}$$

3.8 Confidence

In using bilingual seed words such as 病毒/*virus* as “bridges” for word translation, the quality of the bilingual seed lexicon naturally affects the system output. First, segmentation of the Chinese text into words already introduces some ambiguity of the seed word identities. Secondly, English-Chinese translations are complicated by the fact that the two languages share very little stemming properties, or part-of-speech set, or word order. This property causes every English word to have many Chinese translations and vice versa. In a source-target language translation scenario, the translated text can be “rearranged” and cleaned up by a monolingual language model in the target language. However, the lexicon is not very reliable in establishing “bridges” between non-parallel English-Chinese texts. To compensate for this ambiguity in the seed lexicon, we introduce a **confidence weighting** to each bilingual word pair used as seed words. If a word i_e is the $k-th$ candidate for word i_c, then $w_{i_{t_e}} = w_{i_{t_e}}/k_i$.

The similarity score then becomes:

$$\begin{aligned} S'(W_c, W_e) &= \frac{\Sigma_{i=1}^{t}(w_{ic} \times w_{ie})/k_i}{\sqrt{\Sigma_{i=1}^{t} {w_{ic}}^2 \times \Sigma_{i=1}^{t} {w_{ie}}^2}} \\ \text{where } \ w_{ic} &= TF_{ic} \times IDF_i \\ w_{ie} &= TF_{ie} \times IDF_i \end{aligned}$$

3.9 Experimental Results

Evaluation I: unknown words In order to apply the above algorithm to find the translation for 流感/*liougan* from the newspaper corpus, we first use a script to select the 118 English content words which are not in the lexicon as possible candidates. The highest ranking candidates of 流感 are *flu, Lei, Beijing, poultry* respectively. We also apply the algorithm to the frequent Chinese unknown words and the 118 English unknown words from the English newspaper. The output is ranked by the similarity scores. The highest ranking translated pairs are shown in Table 6.

Evaluation II: known words A second evaluation is carried out on randomly selected 40 known English words from the English newspaper against 900 known Chinese words from the Chinese newspaper. This evaluation is more automatic because a dictionary can be used to find correct translations. We have added *flu/*流感 in the dictionary.

The five highest ranking candidates for *flu, shop, virus* are shown in Table 7.

For the test set of 40 English words against 900 Chinese candidates, translation accuracy ranges from 30% when only the top candidate is counted, to 76% when top 20 candidates are considered, and up to 88% when top 40 are counted. We suggest that it is not unreasonable for the system to to give 20+ translation candidates for each word when the system is used as translator-aid.

4 Discussion of results

Predictably, the characteristics of non-parallel corpora cause lexicon extraction accuracy to be lower than that obtained from parallel corpora of similar sizes. Some other errors are caused by inherent differences between English and Chinese:

1. **Polysemous words**: Online newspapers cover multiple sub-domains. There is a much higher chance for words to be used in multiple senses. In our evaluation set II, such words include *drive, behind, fine, track,* etc. We expect this type of error to reduce when we only translate domain words or terms.
2. **Many-to-many translations**: *spread* has seven Chinese equivalents and *virus* has three in our corpus. Each of the Chinese equivalent in turn has

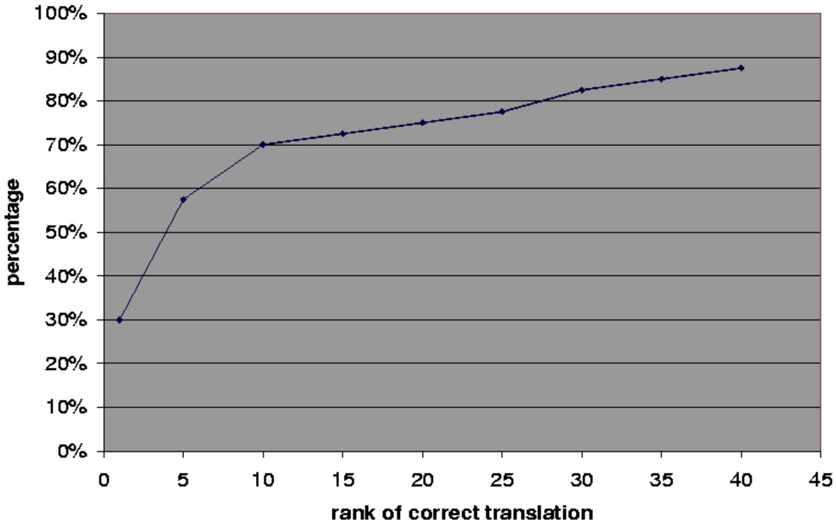
Table 6. Convec output for unknown English and Chinese words

score	English	Chinese
0.008421	Teng-hui	登輝 (Teng-hui)
0.007895	SAR	特區 (SAR)
0.007669	flu	流感 (flu)
0.007588	Lei	鴨 (Lei)
0.007283	poultry	家禽 (Poultry)
0.006812	SAR	建華 (Chee-hwa)
0.006430	hijack	登輝 (Teng-hui)
0.006218	poultry	特區 (SAR)
0.005921	Tung	建華 (Chee-hwa)
0.005527	Diaoyu	登輝 (Teng-hui)
0.005335	PrimeMinister	登輝 (Teng-hui)
0.005335	President	登輝 (Teng-hui)
0.005221	China	林 (Lam)
0.004731	Lien	登輝 (Teng-hui)
0.004470	poultry	建華 (Chee-hwa)
0.004275	China	登輝 (Teng-hui)
0.003878	flu	鴨 (Lei)
0.003859	PrimeMinister	建華 (Chee-hwa)
0.003859	President	建華 (Chee-hwa)
0.003784	poultry	梁 (Leung)
0.003686	Kalkanov	珠海 (Zhuhai)
0.003550	poultry	鴨 (Lei)
0.003519	SAR	葉利欽 (Yeltsin)
0.003481	Zhuhai	建華 (Chee-hwa)
0.003407	PrimeMinister	林 (Lam)
0.003407	President	林 (Lam)
0.003338	flu	家禽 (Poultry)
0.003324	apologise	登輝 (Teng-hui)
0.003250	DPP	登輝 (Teng-hui)
0.003206	Tang	唐 (Tang)
0.003202	Tung	梁 (Leung)
0.003040	Leung	梁 (Leung)
0.003033	China	特區 (SAR)
0.002888	Zhuhai	農曆 (Lunar)
0.002886	Tung	董 (Tung)

Table 7. Top five translation candidates for *flu*, *shop*, *drug*

flu	Chinese	Gloss	shop	Chinese	gloss	drug	Chinese	gloss
0.082867	傳播	spread	0.031571	商店	shop	0.056118	藥	drug
0.064126	傳染	contagious	0.014100	價格	cost	0.019768	昏睡	lethargy
0.056328	感染	contraction	0.012916	算出	figure$_o ut$	0.019208	查出	discover
0.054037	擴散	diffuse	0.011575	施與	send	0.017252	臀部	behind
0.046364	流感	flu	0.011552	飼養	breed	0.015905	病人	invalid

Fig. 3. Translation accuracy of N-best candidates



multiple English counterparts. This splits the context vectors into smaller pieces and gives lower scores to each translations. For example, all three correct translations for *virus* rank lower than another word, *contagious*. But when they are combined into a single context vector, the final score ranks first in the list.

3. **Dictionary error**: The bilingual dictionary we use is augmented by a statistical process. Its content includes many context-dependent entries in addition to human errors. This causes errors in our output as well.
4. **Missing candidate**: It is not surprising that non-parallel, comparable corpora contain a lot of un-matchable words.
5. **Chinese tokenization**: One complication in the translation process of Chinese words is that the words have no space delimiters. The tokenizer we use to insert word boundaries makes mistakes when the words are not found in its dictionary. This reduces the number of context seed words we can use for matching context vectors.
6. **Stemming**: There are mismatches between lexical types in English and Chinese. For example, the Chinese translations for *spread*, *spreading* are all the same, and the Chinese translations for *beauty*, *beautiful* and *beautifully* are often the same. To compensate for such mismatches, it might be helpful to use a stemming tool to pre-process the corpora. However, some of the morphological information might be useful for translation. So it is not clear what the optimal amount of stemming is needed for bilingual lexicon extraction.

Some of these errors can be overcome by using better tools or some learning procedures. Some others, such as the missing candidate problem, is inherent in comparable corpora. To solve this problem, the *size* of the comparable corpus has to be so large as to include all possible words.

5 Conclusion

In this paper, we present our view on the general approaches and the evolution of statistical bilingual lexicon extraction, starting from using aligned parallel corpora as resources, to noisy parallel corpora, and continuing with comparable corpora. Most importantly, we discuss the paradigm change from parallel corpora to comparable corpora and suggest a new method, **Convec**, to find bilingual lexicon in comparable corpora. **Convec** finds translations of new words by matching its context vector with that of its counterpart in the target language. We have tested this method on a comparable corpora consisting of texts from various English and Chinese newspaper articles. Two sets of evaluations are carried out. The first test set consists of all English and Chinese unknown words from the corpus. **Convec** matches most of the unknown English word to their correct translation in Chinese. The second test set consists of 40 English words matched against 900 Chinese words. Translation accuracy is 30% when top one candidate is counted, 76% when top 20 are counted and 88% when top 40 are included. Translation candidates among the top 20 are usually words related to the true translation. This means that one can refine the algorithm eventually to get better precision. We conclude that bilingual lexicons extracted from non-parallel, comparable corpora can be used to augment dictionaries containing a baseline lexicon extracted from parallel corpora, or entered by human lexicographers.

6 Acknowledgement

Students at the Human Language Technology Center, HKUST, Lo Yuen Yee and Luis Parmentier(visiting student from ENST, France), both implemented different parts of the programs for **Convec**, and Liu Yi helped with the evaluations for the output of **Convec**. The newspaper material is downloaded with a software tool developed by Xuen Yin Xia.

References

1. A. Bookstein. Explanation and generalization of vector models in information retrieval. In *Proceedings of the 6th Annual International Conference on Research and Development in Information Retrieval*, pages 118–132, 1983.
2. P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P. Roosin. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85, 1990.
3. P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

4. Stanley Chen. 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, Ohio, June 1993.
5. W. Bruce Croft. A comparison of the cosine correlation and the modified probabilistic model. In *Information Technology*, volume 3, pages 113–114, 1984.
6. Ido Dagan and Kenneth W. Church. Termight: Identifying and translating technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 34–40, Stuttgart, Germany, October 1994.
7. Ido Dagan, Kenneth W. Church, and William A. Gale. 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, Ohio, June 1993.
8. Ido Dagan and Alon Itai. Word sense disambiguation using a second language monolingual corpus. In *Computational Linguistics*, pages 564–596, 1994.
9. Pascale Fung. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the Third Annual Workshop on Very Large Corpora*, pages 173–183, Boston, Massachusettes, June 1995.
10. Pascale Fung. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics*, pages 236–233, Boston, Massachusettes, June 1995.
11. Pascale Fung. Domain word translation by space-frequency analysis of context length histograms. In *Proceedings of ICASSP 96*, volume 1, pages 184–187, Atlanta, Georgia, May 1996.
12. Pascale Fung and Kenneth Church. Kvec: A new approach for aligning parallel texts. In *Proceedings of COLING 94*, pages 1096–1102, Kyoto, Japan, August 1994.
13. Pascale Fung and Kathleen McKeown. 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*, pages 81–88, Columbia, Maryland, October 1994.
14. Pascale Fung and Kathleen McKeown. A technical word and term translation aid using noisy parallel corpora across language groups. *Machine Translation*, pages 53–87, 1996.
15. Pascale Fung and Kathleen McKeown. Finding terminology translations from non-parallel corpora. In *The 5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong, Aug. 1997.
16. Pascale Fung and Lo Yuen Yee. An ir approach for translating new words from nonparallel, comparable texts.
17. W. Gale, K. Church, and D. Yarowsky. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics, 1992.
18. W. Gale, K. Church, and D. Yarowsky. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of TMI 92*, 1992.
19. W. Gale, K. Church, and D. Yarowsky. Work on statistical methods for word sense disambiguation. In *Proceedings of AAAI 92*, 1992.
20. W. Gale, K. Church, and D. Yarowsky. A method for disambiguating word senses in a large corpus. In *Computers and Humanities*, volume 26, pages 415–439, 1993.
21. William Gale and Kenneth Church. Identifying word correspondences in parallel text. In *Proceedings of the Fourth Darpa Workshop on Speech and Natural Language*, Asilomar, 1991.

22. M. Hearst. Noun homograph disambiguation using local context in large text corpora. In *Using Corpora*, Waterloo, Canada, 1991.
23. Martin Kay and Martin Röscheisen. Text-Translation alignment. *Computational Linguistics*, 19(1):121–142, 1993.
24. Robert Korfhage. Some thoughts on similarity measures. In *The SIGIR Forum*, volume 29, page 8, 1995.
25. Julian Kupiec. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, pages 17–22, Columbus, Ohio, June 1993.
26. Frederick Mosteller and David L. Wallace. *Applied Bayesian and Classical Inference - The Case of The Federalist Papers.* Springer Series in Satistics, Springer-Verlag, 1968.
27. G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.
28. Hinrich Shütze. Dimensions of meaning. In *Proceedings of Supercomputing '92*, 1992.
29. Frank Smadja, Kathleen McKeown, and Vasileios Hatzsivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 21(4):1–38, 1996.
30. Howard R. Turtle and W. Bruce Croft. A comparison of text retrieval methods. In *The Computer Journal*, volume 35, pages 279–290, 1992.
31. Dekai Wu and Hongsing Wong. Machine translation with a stochastical grammatical channel.
32. Dekai Wu and Xuanyin Xia. Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 206–213, Columbia, Maryland, October 1994.
33. D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Conference of the Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics, 1995.