# Automatic Extraction of English-Chinese Term Lexicons from Noisy Bilingual Corpora

**Sun Le, Jin Youbing, Du Lin, Sun Yufang**

Open System & Chinese Information Processing Center
Institute of Software
Chinese Academy of Sciences
Beijing 100080
P. R. China

lesun, ybjin, yfsun, ldu@sonata.iscas.ac.cn

**Abstract**

This paper describes our system, which is designed to extract English-Chinese term lexicons from noisy complex bilingual corpora and use them as translation lexicon to check sentence alignment results. The noisy bilingual corpora are aligned firstly by our improved length based statistical approach, which could detect sentence omission and insertion partly. A term extraction system is used to obtain term translation lexicons form roughly aligned corpora. Then the statistical approach is used to align the corpora again. Finally, we filter the noisy bilingual texts and obtain nearly perfect alignment corpora.

## 1. Introduction

One of the main problems in human communication is the presence of a huge variety of languages in the world. With the development of the performance of computers, it is become possible to find ways to support the communication of people from different parts of the world.

In the last few years, there has been a growing interest in multilingual corpora. The advantage of processing a multilingual corpus is to obtain context specific information between these languages, which are usually much less ambiguous than general collections. They have been used in many domains such as automatic or human-aid translation, multilingual terminology and lexicography, multilingual information retrieval systems, etc.

The first step to extract structural information and statistical parameters from multilingual corpora is sentence alignment. This problem has been well studied and a number of quite encouraging results have been reported.

However, the performance tends to deteriorate significantly when these approaches are applied to noisy corpora which are widely different from the training corpus and/or which are less literal translation (with sentence omission or insertion, which are very common in real texts).

In order to increase both the robustness and accuracy of sentence alignment, a good translation term lexicon is needed (with different to different corpora, such as bilingual texts of law, special science domain and literary), especially to English-Chinese bilingual corpora for there are no cognates at all. This motivates the research of our paper.

In the following sections, we first describe related work on sentence alignment and lexicon extraction. This section does not contain a complete survey of all-existing methods and techniques in these research areas, but contains the important approaches with respect to the implementations of our system. After presenting the outline of our algorithm, we describe the method for detecting sentence omission and insertion, and then give the method we used to extract term lexicon. Finally, we present our results and describe directions for future work.

## 2. Related Work

The recent availability of large amount of bilingual corpora has inspired the interest of researchers in several areas, such as machine translation, human-aid translation, multilingual terminology, etc. Our research is related to two areas: sentence alignment and lexicon extraction.

There are basically three kinds of approaches on sentence alignment: the length-based approach (Gale & Church 1991 and Brown et al. 1991), the lexical approach (kay & Roscheisen 1993), and the combination of them (Chen 1993, Wu 1994 and Langlais 1998, etc.).

The first published algorithms for aligning sentences in parallel texts are length-based approach proposed by Gale & Church (1991) and Brown et al. (1991). Based on the observation that short sentences tend to be translated as short sentences and long sentences as long sentences, they calculate the most likely sentence correspondences as a function of the relative length of the candidates. The basic approach of Brown et al. is similar to Gale and Church, but works by comparing sentence length in words rather than characters. While the idea is simple, the models can still be quite effective when used to clear and literal translated corpora. Once the algorithm had accidentally mis-aligned a pair sentence, it tends to be unable to correct itself and get back on track before the end of the paragraph. Use alone, length-based alignment algorithms are therefore neither very robust nor reliable.

Kay & Roscheisen (1993) use a partial alignment of lexical items to induce a maximum likelihood at sentence level. The method is reliable but time consuming.

Chen (1993) combines the length-based approach and lexicon-based approach together. A translation model is used to estimate the cost of a certain alignment, and the best alignment is found by using dynamic programming as the length-based method. The method is robust, fast enough to be practical and more accurate than previous methods.

The first sentence alignment model used to align English-Chinese bilingual texts is proposed by Wu (1994). For lack of cognates in English-Chinese, he used lexical cues to add the robust of his model.

All of these works are test on nearly clear and literal translation bilingual corpora.

There are many projects in corpus linguistics about lexicon extraction, which are based on different corpora and use a variety of different approaches, such as, the Champollion System (Frank Smadja et al. 1993), the Termight System (Dagan & Church 1993), the method for Acquisition of Bilingual Terminology (Pim van der Eijk 1993), the K-vec method (Pascale Fung 1994), and the English-Chines lexicon extraction system (Dekai Wu 1996).

The Champollion System focuses on the identification of collocations and the automatic extraction of corresponding translations in a given parallel bilingual corpus. The Termight System, based on part-of-speech tagging and the word alignment, is a tool for the identification of technical terms and the support of translation processes. The method of Pim van der Eijk (1993) concentrates on identifying noun phrases from a previously aligned and tagged parallel corpus. The K-vec Method is to extract lexicon candidates by looking for similarities in the distribution of source and target language word. Dekai Wu (1995) use an estimation maximization algorithm with additional filter techniques to extracting single word translations from a sentence aligned parallel corpus.

## 3. Outline of the Algorithm

In this section we present the outline of the algorithm in our system, which is designed to extract English-Chinese translation lexicons from noisy parallel corpora and used them as term lexicons to check sentence alignment results. The outline of our algorithm is shown as follows:

Step 1: The sentence boundary of bilingual texts is identified by a heuristic method.

Step 2: A Chinese word segmentation model is introduced to segment the Chinese Characters to words roughly by a Chinese common word lexicon.

Step 3: A primary English-Chinese Lexicon is used to separate the bilingual texts into a few shorter bilingual texts correctly by heuristic search.

Step 4: The separated bilingual texts are aligned respectively by our newly improved statistical algorithm, which is based on the well-known statistical model of character lengths. The trouble of sentence omission and insertion are partly resolved by this algorithm.

Step 5. A lexicon check process is added to judge all the alignment results in last step by the primary English-Chinese lexicon. A score S is given to every alignment sentence pair. The alignments whose score below a certain constant $C_b$ are judged as noisy alignment and removed from the bilingual texts temporally.

Step 6. The rest aligned bilingual texts are used to extract a translation term lexicon by co-occurrence probability and the part of speech of words. It's not a simple task for English-Chinese bilingual corpora because there are always some wrong segment of Chinese word.

Step 7: Repeat Step 4 to align the separated bilingual texts again.

Step 8. The check process is introduced again to judge all the alignment results by the primary English-Chinese lexicon and the newly extracted term lexicon in step 6. The alignments whose score S above a certain constant $C_a$ are judged as correct alignment. Finally, we filter the noisy bilingual texts and obtain nearly correct alignment parallel corpora.

## 4. Omission and Insertion Detection

It's quit obvious that the performance of length based alignment tends to deteriorate significantly when there are sentence omission and/or insertion in bilingual texts. In our experiment just one or two sentence omission in one language can decrease the correct rate greatly. The former statistical algorithm based on character length never gets these kinds of alignment correct. However, a few paragraph or sentence omissions in large-scale bilingual corpora are common in real texts.

In our system, a new improved statistical algorithm is proposed. The key idea of this improved algorithm is the introduction of an assumption that this sentence may be omitted in one language in every step of dynamic programming algorithm. A probabilistic score is given to suggest the likelihood of the omission by character length of this sentence. Then compare that score with the score where the sentence isn't omitted and choose the better. For example, if there are four kinds of possible sentence alignment classes, such as 1:1,2:1,1:2,2:2. Let D( i, j)be the maximum likelihood alignment between sentences $S_1,\ldots,S_i$ and $t_1,\ldots, t_j$.

Then one can recursively define and calculate D( i, j) by using the initial condition D(0,0)=0, and defining:

$$D(i,j)=\min\begin{cases} D(i-1,j-1)+\cos t(s_i;t_j) \\ D(i-1,j-2)+\cos t(s_i;t_{j-1},t_j) \\ D(i-2,j-1)+\cos t(s_{i-1},s_i;t_j) \\ D(i-2,j-2)+\cos t(s_{i-1},s_i;t_{j-1},t_j) \end{cases} \quad (1)$$

Where $\text{cost}(s_i;t_j)$ can be calculated from follow equation accoring to the gaussian assumption. For detail see Gale & Church(1991).

$$\delta=\frac{l_t-l_sC}{\sqrt{l_sS^2}} \quad (2)$$

where $l_s$ and $l_t$ mean the character length of sentence of source language and target language

respectively. C is the mean number of target language characters generated by each source language character.

However, in our system, the cost($s_i$;$t_j$) is defining as follow:

$$\cos t(s_i;t_j)=\min\begin{cases}\cos t(s_i;t_j)\\ \cos t(s_{i-1};t_j)\times C_W\\ \cos t(s_i;t_{j-1})\times C_W\end{cases} \quad (3)$$

Where cost($s_{i-1}$;$t_j$) means the cost of alignment sentence $S_{i-1}$ with $t_j$, while the sentence $S_i$ in source text is omitted during the computation; cost($s_i$;$t_{j-1}$) means the cost of alignment sentence $S_i$ with $t_{j-1}$, while the sentence $t_j$ in target text is omitted; $C_W$ is a constant used to decrease the prior of these kinds of alignment.

Therefore, the minimum cost among alignment ($s_i$;$t_j$), alignment ($s_{i-1}$;$t_j$) and alignment ($s_i$;$t_{j-1}$) is chosen by character length. Similarly, we can get the expression of cost($s_i$; $t_{j-1}$, $t_j$), cost($s_{i-1}$,$s_i$;$t_j$) and cost($s_{i-1}$ ,$s_i$; $t_{j-1}$, $t_j$).By this improved algorithm, some omission and insertion of bilingual texts are identified in our experiment.

## 5. Lexicon Checking

It's obviously difficult to increase greatly the accuracy and robust of sentence alignment only by length based approach. So a lexicon checking process is added to our system. The alignment results obtained by length based approach are checked by a primary English-Chinese lexicon. A score $S_A$ is given to every alignment sentence pair. The alignments whose score below a certain constant $C_b$ are judged as noisy alignment and removed from the bilingual texts temporally. After extracting the term lexicon from the roughly aligned texts we align the whole corpora again. Then we use the newly extracted term lexicon to check the alignments results, whose score S above a certain constant $C_a$ are judged as correct alignment. The score $S_A$ is calculated by simple equation as follow:

$$S_A=\frac{No_{correct}\times 2}{No_{English}+No_{Chinese}} \quad (4)$$

that is, the twice number of correctly matched English words and Chinese words to the sum of total number of English and Chinese words in one aligned sentence pair.

## 6. Term Extraction

As we are interested in finding domain specific terms as term lexicons, we tagged the English part of the corpus using a POS tagger, extracted noun phrases which are more likely to be term. The patterns of term we consider are as follows: N, AN, NN, AAN, NNN, NAN, NPN. In these patterns A refers to an adjective, P to a preposition, and N to a noun.

Then we use a Chinese word POS tagger to tag the Chinese part of the corpus also. It's not a simple task for Chinese texts for there are always some wrong segments for Chinese words and sometime the noun or adjective words are translated to verb or noun words in English. A heuristic filter is used to get away the words, which are not likely to be Chinese term, such as conjunction, pronoun, numeral and other most frequent words. Then we calculates local frequencies (the frequency of the English candidate term in the subset of the alignments containing the Chinese candidate term) and global frequencies of English candidate term and use the following quotient for measuring the correlation.

$$f_{ec}=\frac{f_{local}(English,Chinese)}{f_{global}(English)} \quad (5)$$

Only the English terms that occur in the corpus above 10 times are considered candidate term. Finally we choice the Chinese term whose score is above a specific threshold as the translation of this English candidate term. An example is given in fig. 1. We can see the Chinese words 剑士, 锦衫, which are segmented as 剑 and 士, 锦 and 衫 can be recover.

*English Sentence with POS tag*: The(z) swordsman(n) in(p) blue(n) cut(v) three(m) times(n). The(z) liveried(a) swordsman(n) blocked(v) each(r) cut(n).(w)
*Chinese Sentence with POS tag*: 青衣(n) 剑(n) 士(n) 连劈(v) 三(m) 剑(n)，锦(n) 衫(n) 剑(n) 士(n) 一一(d) 格(n) 开(v) 。(w)
*English Candidate terms*: liveried swordsman; swordsman in blue; swordsman
*Chinese Candidate terms:* 剑; 士; 锦; 衫; 青衣
*Correlation Score:*

| liveried swordsman | | swordsman in blue | | swordsman | |
|---|---|---|---|---|---|
| 1.00 | 剑 | 1.00 | 剑 | 0.95 | 剑 |
| 0.84 | 士 | 0.83 | 士 | 0.85 | 士 |
| 0.80 | 锦 | 0.88 | 青衣 | 0.51 | 青衣 |
| 0.80 | 衫 | 0.42 | 锦 | 0.51 | 锦 |
| 0.40 | 青衣 | 0.38 | 衫 | 0.50 | 衫 |

Figure 1. An example of term extraction

## 7. Experiment Results

We tested our system with two different style English-Chinese bilingual corpora that correspond to different domain. One is the story ('sword of the Yuen Maiden', author, Jin Yong), and the other is computer handbook (Sco Unix handbook). There are about 843 English sentence and 694 Chinese sentence in the story and 3274 English sentence and 3186 Chinese sentence in the computer handbook. Finally we extracted 24 terms and 61 terms from the two corpora and use them with the primary English-Chinese translation lexicon together as lexicon to check the noisy bilingual corpora. Four Examples of unfiltered term with score $f_{ec}$ are given in table 1. No matter the Chinese segment is right or not, the translation of term can be find. Some examples of sentence alignment results are showed in figure 2. The omission is correctly identified by our system. The detail experiment results of sentence alignment after Step 4 to computer handbook are given in table 2. Finally we introduce the term lexicon and primary lexicon to check them and obtain nearly perfect results.

## 8. Conclusion and Future Work

Although sentence alignment usually is a manageable problem, there are situations where even humans have hard time making the right decision. The performance is directly related with the complexity of bilingual corpora used in test. The trouble of translation omission and insertion are notorious to cope with.

This paper describes our system, which is designed to extract English-Chinese term lexicons from noisy complex bilingual corpora and use them as lexicon to check sentence alignment results. The noisy bilingual corpora are aligned by our improved length based statistical approach, which could detect sentence omission and insertion partly. Finally, we filter the noisy bilingual texts and obtain nearly perfect alignment corpora.

Although the results we got are quite promising to complex bilingual texts, there are still much to do in near future, Such as, to increase the correct rate of Chinese word segmentation, to add a synonym lexicon during match of English words with Chinese words, etc.

## 9. References


Brown P. F., Lai, J. C., and Mercer, R. L., 1991. *Aligning Sentences in Parallel Corpora.* In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), pp.169-176.

Frank Smadja, 1993 *Retrieving Collocations from Text: XTRACT. Computational Linguistics*

Fung, P., and Church, K. W., 1994. *K-vec: A New Approach for Aligning Parallel Texts.* In Proceedings of the 15th International Conference on Computational Linguistics (COLING'94), Tokyo, Japan, pp. 1096-1102,

Gale, W. A., and Church, K. W., 1991. *A Program for Aligning Sentences in Bilingual Corpora.* In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), pp. 177-184

I. D. Melamed. 1996. *Automatic Detection of Omissions in Translations.* In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark

J.S. Chang and M. H. Chen, 1997. *An alignment method for noisy parallel corpora based on image processing techniques.* In Proceedings of the 35th Meeting of the Association for Computational Linguistics, Madrid, pp. 297-304

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

Ph. Langlais, M. Simard, J. Veronis, S.Armstong, P. Bonhomme, F. Debili, P. Isabelle, E. Souissi, and P. Theron, 1998. *Arcade: A cooperative research project on parallel text alignment evaluation*. In First International Conference on Language Resources and Evaluation, Granada, Spain.

Wu Daikai., 1994. *Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria.* In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94), pp.80-87

Wu Daikai.and Xia Xuanyin. 1995. *Large-Scale Automatic Extraction of an English-Chinese translation Lexicon.* Machine Translation, 9:3-4, 285-313

| boot disk | floppy disk | device driver | hardware self-check stops |
|---|---|---|---|
| 1.00 盘 | 1.00 软 | 1.00 设备 | 1.00 硬件 |
| 0.89 引导 | 1.00 盘 | 1.00 驱动 | 1.00 自 |
| 0.22 机器 | 0.33 引导 | 1.00 程序 | 1.00 检 |
| 0.16 软 | 0.33 插入 | 0.28 过程 | 1.00 停止 |
| 0.11 替换 | 0.20 门 | 0.27 停止 | 0.50 系统 |
| 0.06 制造 | 0.07 正常 | 0.14 初始 | 0.38 安装 |
| 0.05 检验 | 0.06 程序 | 0.03 问题 | 0.24 错误 |

Table 1. Examples of unfiltered output with score

Decision = 2-1 S_Char = 235 T_Char = 165
The Release Notes, SCO OpenServer Handbook, and SCO OpenServer Internet Services are provided in printed format with every SCO OpenServer system package.
The SCO OpenServer Handbook is also available online, along with many other books.
《Release Notes》及《SCO OpenServer Handbook》以出版的文档形式随 SCO OpenServer 系统软件包一起提供，同时《SCO OpenServer Handbook》及其它一些手册还以联机手册形式提供。

Decision = 3-1 S_Char = 208 T_Char = 124
Most of the online books are available in printed format from your vendor, in two sets.
Set 1 includes all the user's and administrator's guides.
Set 2 includes all the reference manual pages (several volumes).
大多数出版的联机手册可以从销售商处得到，手册分两部分，一部分包括所有的用户与管理中指南，另一部分包括所有的参考手册(若干卷)。

Decision = 1-0 S_Char = 94 T_Char = 0
See Related documentation for detailed information about the SCO OpenServer documentation set.

Decision = 1-1 S_Char = 32 T_Char = 22
Additional licenses and products
附加的许可证及产品

Figure 2. Four examples of sentence alignment results

| Class of Alignment | No. of Aligned Sentence Pair | No. of Correct Sentence Pair | No. of Error Sentence Pair | Precision after Step 4 |
|---|---|---|---|---|
| 1:1 | 2096 | 2071 | 25 | 98.81% |
| 1:2 | 291 | 261 | 30 | 89.69% |
| 2:1 | 241 | 223 | 18 | 92.53% |
| 2:2 | 63 | 56 | 7 | 88.89% |
| 1:3 | 36 | 27 | 9 | 75.00% |
| 3:1 | 33 | 26 | 7 | 78.79% |
| 2:3 | 3 | 2 | 1 | 66.67% |
| 3:2 | 3 | 1 | 2 | 33.33% |
| 3:3 | 2 | 1 | 1 | 50.00% |
| 0:1 | 4 | 2 | 2 | 50.00% |
| 1:0 | 6 | 4 | 2 | 66.67% |
| Total | 2778 | 2674 | 104 | 96.26% |

Table 2: The detail experiment results to our test