Word Alignment of English-Chinese Bilingual Corpus Based on Chunks

Sun Le, Jin Youbing, Du Lin, Sun Yufang
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

In this paper, a method for the word alignment of English-Chinese corpus based on chunks is proposed. The chunks of English sentences are identified firstly. Then the chunk boundaries of Chinese sentences are predicted by the translations of English chunks and heuristic information. The ambiguities of Chinese chunk boundaries are resolved by the coterminous words in English chunks. With the chunk aligned bilingual corpus, a translation relation probability is proposed to align words. Finally, we evaluate our system by real corpus and present the experiment results.

Key Words: Word Alignment, Chunk Alignment, Bilingual Corpus, Lexicon Extraction

1 Introduction

With the easier access to bilingual corpora, there is a tendency in NLP community to process and refine the bilingual corpora, which can serve as the knowledge base in support of many NLP applications, such as automatic or human-aid translation, multilingual terminology and lexicography, multilingual information retrieval system, etc.

Different NLP applications need different bilingual corpora, which are aligned at different level. They can be divided by the nature of the segment to section level, paragraph level, sentence level, phrase level, word level, byte level, etc.

As for our applications, we choose the chunk level to do alignment based on following considerations. Firstly, our applications, which include an example-based machine translation system, a computer aid translation system and a multilingual information retrieval system, need the alignment below the sentence level, on which we can acquire bilingual word and phrase dictionaries and other useful translation information. Secondly, the word level alignment between English and Chinese language is difficult to deal with. There are no cognate words. The change in Chinese word order and word POS always produce many null and mistake correspondences. Next, we observe the phenomenon that when we translate the English sentence to Chinese sentence, all the words in one English chunk tend to be translated as one block of Chinese words which are coterminous. The word orders within these blocks tend to keep with the English chunk, also. So there are stronger boundaries between chunks than between words when we translate texts. Finally, as we all known, chunk has been assigned syntactic structure (Steven Abney, 1991), which comprises a connected sub-graph of the sentence's parse tree. So it's possible to align sentence structure and obtain translation grammars based on chunks by parsing.

Many researchers have studied the text alignment problem and a number of quite encouraging results have been reported to different level alignments. With sentence-aligned corpus ready in hand, we focus our attention on the intra-sentence alignment between the sentence pairs. In this paper, a method for the word alignment of English-Chinese corpus based on chunks is proposed. The chunks of English sentences are identified firstly. Then the chunk boundaries of Chinese sentences are predicted by the bilingual lexicon and synonymy Chinese dictionary and heuristic information. The ambiguities of Chinese chunk boundaries are resolved by the coterminous words in English chunks. With the chunk aligned bilingual corpus, a translation relation probability is proposed to align words. Although this paper is related to English-Chinese word alignment, the idea can be used to any other language bilingual corpora. In the following sections, we first present a brief review of related work in word alignment. Then discuss our alignment algorithm based on chunks in detail. Following this is an analysis of our experimental results. Finally, we close our paper with a discussion of future work.

2 Related Work

There are basically two kinds of approaches on word alignment: the statistical-based approaches (Brown et. al., 1990; Gale & Church, 1991; Dagan et. al. 1993; Chang, 1994), and the lexicon-based approaches (Ker & Chang, 1997; Wang ét. al., 1999).

Several translation models based on word alignment are built by Brown et al. (1990) in order to implement the English-French statistical machine translation. The probabilities, such as translation probability, fertility probability, distortion probability, are estimated by EM algorithm. The χ^2 measure is used by Gale & Church (1991) to align partial words. Dagan (1993) uses an improved Brown model to align the words for texts including OCR noise. They first align word partially by character string matching. Then use the translation model to align words. Chang (1994) uses the POS probability rather than translation probability in Brown model to align the English-Chinese POS tagged corpus. Ker & Chang (1997) propose an approach to align Chinese English corpus based on semantic class. There are two semantic classes are used in their model. One is the semantic class of Longman lexicon of contemporary English, the other is synonymy Chinese dictionary. The semantic class rules of translation between Chinese and English are extracted from large-scale training corpus. Then Chinese and English words are aligned by these rules. Wang (1999) also uses the lexicons to align the Chinese English bilingual corpus. His model is based on bilingual lexicon, sense similarity and location distortion probability.

The statistical-based approaches need complex training and are sensitive to training data. It's a pity that almost no linguistic knowledge is used in these approaches. The lexicon-based approaches seem simplify the word alignment problem and can't obtain much translation information above word level. To combine these two approaches in a better way is the direction in near future. In this paper we proposed a method to align the bilingual corpus base on chunks. The linguistic knowledge such as POS tag and Chunk tag are used in a simply statistical model.

3 Alignment Algorithm

3.1 Outline of Algorithm

For our procedure in this paper, the bilingual corpus has been aligned at the sentence level, and the English language texts have been tagged with POS tag, and the Chinese language texts have been segmented and tagged with POS tag.

We have available a bilingual lexicon which lists typical translation for many of the words in the corpus. We have available a synonymy Chinese dictionary, also. We identify the chunks of English sentences and then predict the chunk boundaries of Chinese sentences from the translation of every English chunks and heuristic information by use of the bilingual lexicon. The ambiguities of Chinese chunk boundaries are resolved by the coterminous words in English chunks. After produce the word candidate sets by statistical method, we calculate the translation relation probability between every word pair and select the best alignment forms. The detail algorithm for word alignment is given in table 1.

Step 1: According to the definition of Chunk in English, separate the English sentence into a few chunks and labeled with order number from left to right.

Step 2: Try to find the Chinese translation of every English chunk created in step 1 by bilingual dictionary and synonymy Chinese dictionary. If the Chinese translation is find, then label the Chinese words with the same number used for the English chunk in step 1.

Step 3: Disambiguate the multi-label Chinese words by the translation location of coterminous words within the same English chunk.

Step 4: Separate the Chinese sentence into a few chunks by heuristic information.

Step 5: Save all the alignment at chunk level in whole corpus as a base for word alignment.

Step 6: Produce the word candidate sets by statistical method.

Step 7: Calculate the translation relation probability between every word and it's candidate translation words.

Step 8: Select the best translation by comparing the total TRP value in different alignment forms.

Table 1. Outline of Alignment Algorithm

3.2 Chunk Identifying of English Sentence

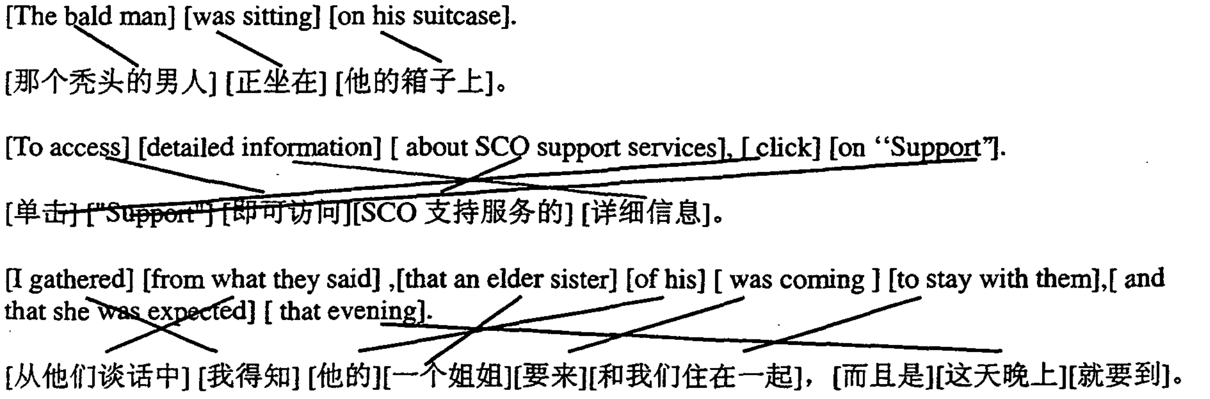
Following Steven Abney (1991), there are two separate stages in chunking parser, which is the chunker and the attacher. The chunker converts a stream of words into a stream of chunks, and the attacher converts the stream of chunks into a stream of sentences. So only the chunker is needed in this paper. It's a non-deterministic version of a LR parser. For detail about chunker and the used grammars, please see Abney (1991). Then the chunks in one sentence are labeled with order number from left to right.

3.3 Chunk Boundary Prediction of Chinese Sentence

We observe the phenomenon that when we translate the English sentence to Chinese sentence, all the words in one English chunk tend to be translated as one block of Chinese words that are coterminous. The word orders within these blocks tend to keep with the English chunk, also. There are three examples in figure 1. The first sentence pair is chosen from an example sentence of Abney (1991). The second sentence pair is from a computer handbook. In these sentence pair all English chunks can find the exactly Chinese Chunk. In the third sentence pair only one English chunk can't find the exactly Chinese chunk for this sentence is chosen from a story and the translation is not literally.

In order to find the Chinese translation of every English chunk, we use the bilingual dictionary and synonymy Chinese dictionary to implement the matching. If the Chinese translation of any words within the English chunk is found, then label the Chinese word with the same number used for labeling the English chunk.

If there are Chinese words, which are labeled simultaneously by two or more number of English chunks, we use the number of nearby Chinese words to disambiguate. For example, in figure 2, the first Chinese word 应用 may be correspondent to the English chunk 5 or 7. We have known that the words in one English chunk tend to be translated as one block of Chinese words that are coterminous. So it's easy to decide the first Chinese word 应用 is correspondent to the English chunk 7, the second Chinese word 应用 is correspondent to the English chunk 5. By the same way, we can find the correct translations of Chinese word 只需 and 需要 is English chunk 6 and chunk 8 respectively. In Step 4 of figure 2, the Chinese words with the same label number are bracketed with in one chunk. Finally, we separate the Chinese sentence into a few chunks by heuristic information based on POS tag (especially the preposition, conjunction, and auxiliary words) and the grammatical knowledge-base of contemporary Chinese (Yu shi wen, 1998).



[The bald man] [was sitting] [on his suitcase].

[那个秃头的男人] [正坐在] [他的箱子上]。

[To access] [detailed information] [about SCO support services], [click] [on “Support"].

[单击] ["Support"] [即可访问][SCO 支持服务的] [详细信息]。

[I gathered] [from what they said] ,[that an elder sister] [of his] [was coming] [to stay with them],[and that she was expected] [that evening].

[从他们谈话中] [我得知] [他的][一个姐姐][要来][和我们住在一起]，[而且是][这天晚上][就要到]。

Figure 1. Three Examples of Chunk Alignment

Step 1 English chunks with order number

[This product 1] [is designed 2] for [low-cost 3], [turnkey solutions 4] and [mission-critical applications 5] that [require 6] [a central application host 7] and [do not require 8] [networking 9].

Step 2 Label the translation of English chunk with it's order number

该(1) 产品(1) 是 为 只需(6/8) 一台(7) 中心(7) 应用(5/7) 主机(7) 而 不(8) 需要(6/8) 联网的(9) 低(3) 成本(3)、高 可靠性的 解决(4) 方案(4) 及 关键性(5) 任务 应用(5/7) 而 设计的(2)。

Step 3 Disambiguate the multi-label Chinese words

该(1) 产品(1) 是 为 只需(6) 一台(7) 中心(7) 应用(7) 主机(7) 而 不(8) 需要(8) 联网的(9) 低(3) 成本(3)、高 可靠性的 解决(4) 方案(4) 及 关键性(5) 任务 应用(5) 而 设计的(2)。

Step 4. Separate the Chinese sentence into a few chunks

[该 产品(1)] 是 为 [只需(6)] [一台 中心 应用 主机(7)] 而 [不 需要(8)] [联网的(9)] [低 成本(3)]、[高 可靠性的 解决 方案(4)] 及 [关键性 任务 应用(5)] 而 [设计的(2)]。

Figure 2. An Example for Chunk Alignment Algorithm from Step 1 to 4

3.4 Calculation of Translation Relation Probability for Words

With the alignments at chunk level of whole corpus, we propose a Translation Relation Probability (TRP) to implement the word alignment. The translation Relation probability of words are given by following equation:

$$P_{ec} = \frac{f_{ec}^2}{f_e \cdot f_c} \qquad (1)$$

Where f_e is the frequency of English word in whole corpus; f_c is the frequency of Chinese Word in whole corpus; f_{ec} is calculated by follow equation:

$$f_{ec} = \sum_{i=1}^{N} \sqrt{\frac{\ln(\frac{2L_{AV}}{L_{ei} + L_{ci}}) + \ln(L_{AV})}{\ln(L_{AV})}} \times \beta_{ec} \qquad (2)$$

Where L_{AV} is the average words number of all English chunks and all Chinese chunks which are related to the English word in whole Corpus; L_{ei} is the word number of the English chunk in which the English candidate words co-occur with the Chinese words; L_{ci} is the word number of the Chinese chunk in which the English candidate words co-occur with the Chinese words; N is the total number of chunks in which the English word co-occur with the Chinese word; β_{ce} is the penalty value to indicate the POS change between the English word and the Chinese word.

By this equation we connect the chunk length and POS change with the co-occurrence frequency. The less the chunk length, the higher the translation relation probability. For example, the chunk pair, which is composed by one English word and two Chinese words, is more reliable than the chunk pair, which is composed by four English words and four Chinese words.

An example is given in figure 3. There are 5 possible alignment forms in our consideration for this chunk, which includes three English words and three Chinese words. Then calculate the total TRP value for every possible alignment word pairs in each alignment form by equation (1). After we get the total TRP value for each alignment form, we choose the biggest one.

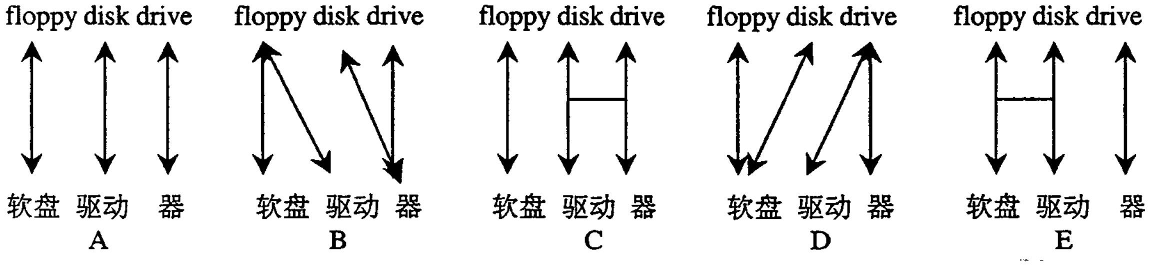


Figure 3. The Possible Word Alignment Forms in One Chunk

4 Experimental Results

4.1 System Architecture

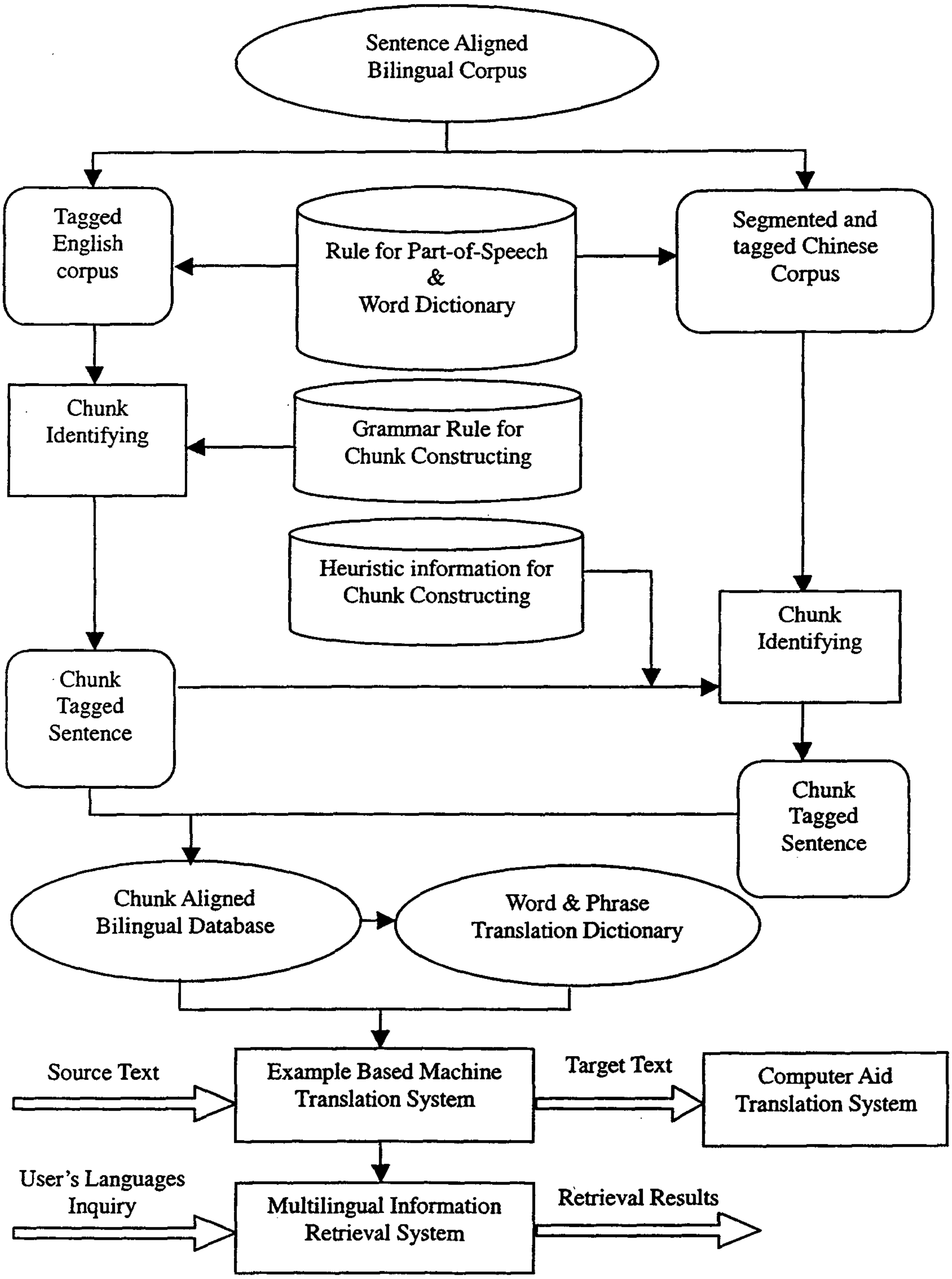

Sentence Aligned Bilingual Corpus
Tagged English corpus
Rule for Part-of-Speech & Word Dictionary
Segmented and tagged Chinese Corpus
Chunk Identifying
Grammar Rule for Chunk Constructing
Heuristic information for Chunk Constructing
Chunk Identifying
Chunk Tagged Sentence
Chunk Tagged Sentence
Chunk Aligned Bilingual Database
Word & Phrase Translation Dictionary
Source Text
Example Based Machine Translation System
Target Text
Computer Aid Translation System
User's Languages Inquiry
Multilingual Information Retrieval System
Retrieval Results

Figure 4. System Architecture

4.2 Experiment Results

We tested our system with an English-Chinese bilingual corpus, which is part of a computer handbook (Sco Unix handbook). There are about 2246 English sentence and 2169 Chinese sentence in this computer handbook after filter noisy figures and tables. Finally we extracted 14,214 chunk pairs from the corpus. The accuracy for automatic chunk alignment is 85.7%. The accuracy for word alignment based on correctly aligned chunk pairs is 93.6%. The errors mainly due to the following reasons: Chinese segmentation error, stop words noise, POS tag error. The parameter β_{ec} we used in equation (2) should be chosen from the training corpus. In table 2, the total TRP values of example in figure 3 are showed. The alignment form D is the best.

(floppy | 软盘)	0.9444 X 1/3	Total TRP of A =0.3792
(disk | 驱动)	0.0212 X 1/3	
(drive | 器)	0.1722 X 1/3	
(floppy | 软盘 驱动)	0.2857 X 1/2	Total TRP of B =0.3194
(disk drive | 器)	0.1765 X 1/2	
(floppy | 软盘)	0.9444 X 1/2	Total TRP of C =0.6485
(disk drive | 驱动 器)	0.3529 X 1/2	
(floppy disk | 软盘)	0.8333 X 1/2	Total TRP of D =0.8640
(drive | 驱动 器)	0.8947 X 1/2	
(floppy disk | 软盘 驱动)	0.3429 X 1/2	Total TRP of E =0.2576
(drive | 器)	0.1722 X 1/2	

Table 2. Total TRP Value for Example in Figure 3

5 Conclusions and Future Work

With the more and more bilingual corpora, there is a tendency in NLP community to process and refine the bilingual corpora, which can serve as the knowledge base in support of many NLP applications. In this paper, a method for the word alignment of English-Chinese corpus based on chunks is presented. After identified the chunks of English sentences, we predict the chunk boundaries of Chinese sentences by the bilingual lexicon, synonymy Chinese dictionary and heuristic information. The ambiguities of Chinese chunk boundaries are resolved by the coterminous words in English chunks. After produce the word candidate sets by statistical method, we calculate the translation relation probability between every word pair and select the best alignment forms. We evaluate our system by real corpus and present the results.

Although the results we got are quite promising to bilingual English Chinese text, there are still much to do in near future. The corpus we use in our experiment is a relative small corpus about computer handbook, in which the terms are translated with high consistency. We should extend our method to the large corpus of other domains without lost much accuracy. To increase the correct rate of Chinese word segmentation is important for our word alignment. To extract the corresponding syntax information of English Chinese bilingual corpus by shallow parsing is a direction for future work, also.

Acknowledgements

This research was. funded by Natural Science Foundation of China (Grant No. 69983009).

The authors would like to thank the anonymous reviewers for their helpful comments.

References

Abney, Steven, 1991. *Parsing by Chunks.* In: Robert Berwick, Steven Abney and Carol Tenny (eds.), Pringciple-Based Parsing, Kluwer Academic Publishers

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

Chang, J. S., and Chen, M. H. C. 1994 *Using Partial*

Aligned Parallel Text and Part-of-speech Information in Word Alignment. In Proceedings of the First Conference of the Association for Machine Translation in the Americas(AMTA'94), pp 16-23

Dagan, I. and Church, K. W. 1994 *Termight: Identifying and Translating.* Technical terminology. In Proceedings of EACL

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), 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

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

Ker, M. and Chang, J. S. 1997 *A Class-Based Approach to Word Alignment.* Computational Linguistics,23(2),pp 313-343

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

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

Sun, Le, Du, Lin, Sun, Yufang, Jin, Youbin 1999 *Sentence Alignment of English-Chinese Complex Bilingual Corpora.* Proceeding of the workshop MAL'99, 135-139

Wang, Bin, Liu, Qun, and Zhang, Xiang, 1999 *An Automatic Chinese-English Word Alignment System.* Proceedings of ICMI'99, pp100-104, Hong Kong

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

Yu, Shiwen, Zhu, Xuefeng, Wang, Hui, Zhang Yunyun, 1998 *The Grammatical Knowledge-base of Contemporary Chinese: A complete Specification.* Tsinghua University Publishers