Multi-level Bootstrapping for Extracting Parallel Sentences from a Quasi-Comparable Corpus

Pascale Fung and Percy Cheung
Human Language Technology Center,
Department of Electrical & Electronic Engineering, HKUST,
Clear Water Bay, Hong Kong
{pascale,eepercy}@ee.ust.hk

Abstract

We propose a completely unsupervised method for mining parallel sentences from *quasi-comparable* bilingual texts which have very different sizes, and which include both in-topic and off-topic documents. We discuss and analyze different bilingual corpora with various levels of comparability. We propose that while better document matching leads to better parallel sentence extraction, better sentence matching also leads to better document matching. Based on this, we use multi-level bootstrapping to improve the alignments between documents, sentences, and bilingual word pairs, iteratively. Our method is the first method that does not rely on any supervised training data, such as a sentence-aligned corpus, or temporal information, such as the publishing date of a news article. It is validated by experimental results that show a 23% improvement over a method without multilevel bootstrapping.

1 Introduction

Sentence-aligned parallel corpus is an important resource for empirical natural language tasks such as statistical machine translation and cross-lingual information retrieval. Recent work has shown that even parallel sentences extracted from comparable corpora helps improve machine translation qualities (Munteanu and Marcu, 2004). Many different methods have been previously proposed to mine parallel sentences from multilingual corpora. Many of these algorithms are described in detail in (Manning and Schűtze, 1999, Dale et al., 2000, Veronis 2001). The challenge of these tasks varies according to the degree of comparability of the input multilingual documents. Existing work extract parallel sentences from parallel, noisy parallel or comparable corpora based on the assumption that parallel sentences should be similar in sentence length, sentence order and bi-lexical context. In our work, we try to find parallel sentences from a *quasi-comparable* corpus, and we find that many of assumptions in previous work are no longer applicable in this case. Alternatively, we propose an effective, multi-level bootstrapping approach to accomplish this task (Figure 1).

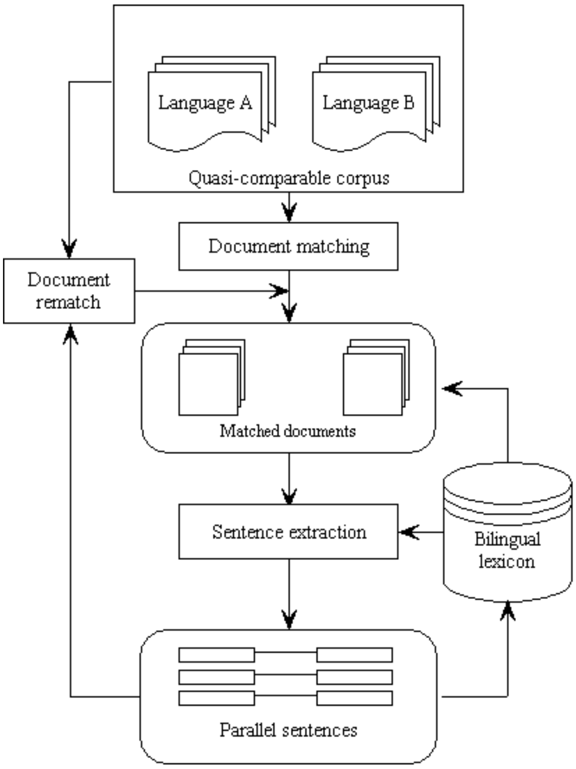

Figure1. Multi-level bootstrapping for parallel sentence extraction

Extraction of matching bilingual segments from non-parallel data has remained a challenging task after almost a decade. Previously, the author and other researchers had suggested that bi-lexical information based on context can still be used to find correspondences between passages, sentences, or words, in non-parallel, comparable texts of the same topic (Fung and McKeown 1995, Rapp 1995, Grefenstette 1998, Fung and Lo 1998, Kikui 1999). More recent works on parallel sentence extraction from comparable data align documents first, before extracting sentences from the aligned documents

(Munteanu and Marcu, 2002, Zhao and Vogel, 2002). Both work used a translation model trained from parallel corpus and adaptively extract more parallel sentences and bilingual lexicon in the comparable corpus. In Zhao and Vogel (2002), the comparable corpus consists of Chinese and English versions of new stories from the Xinhua News agency. Munteanu and Marcu (2002) used unaligned segments from the French-English Hansard corpus and finds parallel sentences among them. Zhao and Vogel (2002) used a generative statistical machine translation alignment model, Munteanu and Marcu (2002) used suffix trees-based alignment model, and Munteanu and Marcu (2004) used a maximum entropy based classifier trained from parallel corpus to extract matching sentences from a comparable corpus of Arabic and English news. The comparable corpora used in all these work consist of documents on the same topic. Our challenge is to find matching bilingual sentences from documents that might or might not be on the same topic.

2 Bilingual Sentence Alignment

There have been various definitions of the term "parallel corpora" in the research community. In this paper, we compare and analyze different bilingual corpora, ranging from the parallel, noisy parallel, comparable, to quasi-comparable.

A *parallel corpus* is a sentence-aligned corpus containing bilingual translations of the same document. The Hong Kong Laws Corpus is a parallel corpus with sentence level alignment; and is used as a parallel sentence resource for statistical machine translation systems. There are 313,659 sentence pairs in Chinese and English. Alignment of parallel sentences from this type of database has been the focus of research throughout the last decade and can be accomplished by many off-the-shelf, publicly available alignment tools.

A *noisy parallel and comparable corpus* contains non-aligned sentences that are nevertheless mostly bilingual translations of the same document. Previous works have extracted bilingual word senses, lexicon and parallel sentence pairs from noisy parallel corpora (Fung and McKeown 1995, Fung and Lo 1998). Corpora such as the Hong Kong News are in fact rough translations of each other, focused on the same thematic topics, with some insertions and deletions of paragraphs.

Another type of comparable corpus is one that contains non-sentence-aligned, non-translated bilingual documents that are topic-aligned. For example, newspaper articles from two sources in different languages, within the same window of published dates, can constitute a comparable corpus. Note that many existing algorithms for sentence alignment from comparable corpus are, in fact, methods for noisy parallel corpus.

On the other hand, a *quasi-comparable corpus* is one that contains non-aligned, and non-translated bilingual documents that could either be on the same topic (*in-topic*) or not (*off-topic*). TDT3 Corpus is a good source of truly non-parallel and quasi-comparable corpus. It contains transcriptions of various news stories from radio broadcasting or TV news report from 1998-2000 in English and Chinese. In this corpus, there are about 7,500 Chinese and 12,400 English documents, covering more than 60 different topics. Among these, 1,200 Chinese and 4,500 English documents are manually marked as being in-topic. The remaining documents are marked as off-topic as they are either only weakly relevant to a topic or irrelevant to all topics in the existing documents. From the in-topic documents, most are found to be comparable. A few of the Chinese and English passages are almost translations of each other. Nevertheless, the existence of considerable amount of off-topic document gives rise to more variety of sentences in terms of content and structure. Overall, the TDT 3 corpus contains 110,000 Chinese sentences and 290,000 English sentences. A very small number of the bilingual sentences are translations of each other, while some others are bilingual paraphrases. In this paper, we describe a method to extract translated and paraphrased bilingual sentence pairs from this quasi-comparable corpus.

2.1 Comparing bilingual corpora

We explore the usability of different bilingual corpora for the purpose of multilingual natural language processing. We argue that the usability of bilingual corpus depends how well the sentences are aligned. To quantify this corpus characteristic, we propose using a lexical alignment score computed from the bilingual word pairs distributed throughout the bilingual sentence pairs.

We first identify bilingual word pairs that appear in the aligned sentence pairs by using a bilingual lexicon (bilexicon). Lexical alignment score is then defined as the sum of the mutual information score of all word pairs that appear in the corpus:

$$S(W_c, W_e) = \frac{f(W_c, W_e)}{f(W_c)f(W_e)}$$

$$S = \sum_{all(W_c, W_e)} S(W_c, W_e)$$

where $f(W_c,W_e)$ is the co-occurrence frequency of bilexicon pair (W_c,W_e) in the aligned sentence pairs. $f(W_c)$ and $f(W_e)$ are the occurrence frequencies of Chinese word *Wc* and English word *We*, in the bilingual corpus.

Table 1 shows the lexical alignment scores of parallel sentences extracted from a parallel corpus (Hong Kong Law), a comparable noisy parallel corpus (Hong Kong News), and a non-parallel, quasi-comparable corpus (TDT 3). We can see that the scores are in direct proportion to the parallel-ness or comparability of the corpus.

Corpus	Parallel	Comparable	Quasi-Comparable
Bilexicon score	359.1	253.8	160.3

Table 1: Bilingual lexicon scores of different corpora

2.2 Comparing alignment assumptions

All previous work on sentence alignment from parallel corpus makes use of one or multiple of the following nine (albeit imperfect) assumptions, as described in the literature (Somers 2001, Manning & Schűtze, 1999), and summarized as below:

1. There are no missing translations in the target document;
2. Sentence lengths: a bilingual sentence pair are similarly long in the two languages;
3. Sentence position: Sentences are assumed to correspond to those roughly at the same position in the other language.
4. Bi-lexical context: A pair of bilingual sentences which contain more words that are translations of each other tend to be translations themselves.

For noisy parallel corpora without sentence delimiters, assumptions made previously for bilingual word pairs are as follows:

5. Occurrence frequencies of bilingual word pairs are similar;
6. The positions of bilingual word pairs are similar;
7. Words have one sense per corpus;
8. Following 7, words have a single translation per corpus;
9. Following 4, the sentence contexts in two languages of a bilingual word pair are similar.

Different sentence alignment algorithms based on both sentence and lexical information can be found in Manning and Schűtze (1999), Wu (2000), Dale et al. (2001), Veronis (2002), and Somers (2002). These methods have also been applied recently in a sentence alignment shared task at NAACL 2003[1]. We have learned that as bilingual corpora become less parallel, it is better to rely on information about word translations rather than sentence length and position.

For comparable corpora, previous bilingual sentence or word pair extraction works are based soly on bilexical context assumption (Fung & McKeown 1995, Rapp 1995, Grefenstette 1998, Fung and Lo 1998, Kikui 1999, Barzilay and Elhadad 2003, Masao and Hitoshi 2003, Kenji and Hideki 2002). Similarly, for quasi-comparable corpora, we cannot rely on any other sentence level or word level statistics but the bi-lexical context assumption. We also postulate one additional assumption:

10. Seed parallel sentences: Documents and passages that are found to contain at least one pair of parallel sentences are likely to contain more parallel sentences.

3 Our approach: Multi-level Bootstrapping

Existing algorithms (Zhao and Vogel, 2002, Munteanu and Marcu, 2002) for extracting parallel sentences from comparable documents seem to follow the 2 steps: (1) extract comparable documents (2) extract parallel corpus from comparable documents. Other work on monolingual, comparable sentence alignment by (Barzilay and Elhadad 2003) also supports that it is advantageous to first align comparable passages and then align the bilingual sentences within the aligned passages. The algorithms proposed by Zhao and Vogel, and by Munteanu and Marcu differ in the training and computation of document similarity scores and sentence similarity scores. Examples of document similarity computation include counting word overlap and cosine similarity. Examples of sentence similarity computation include word overlap count, cosine similarity, and classification scores of a binary classifier trained from parallel corpora, generative alignment classifier. In our work, we use simple cosine similarity measures and we dispense with using parallel corpora to train an alignment classifier. In addition, we do not make any

[1] http://www.cs.unt.edu/~rada/wpt/

```
1. Extract comparable documents
     For all documents in the comparable corpus D:
       a. Gloss Chinese documents using the bilingual lexicon (Bilex);
       b. For every pair of glossed Chinese and English documents, compute document similarity
            =>S(i,j);
       c. Obtain all matched bilingual document pairs whose S(i,j)> threshold1=>C
2. Extract parallel sentences
   For each document pair in C:
       a. For every pair of glossed Chinese sentence and English sentence, compute sentence similarity
            =>S2(i,j);
       b. Obtain all matched bilingual sentence pairs whose S2(i,j)> threshold2=>C2
3. Update bilingual lexicon with unknown word translations
     For each bilingual word pair in C2;
       a. Compute correlation scores of all bilingual word pairs =>S3(i,j);
       b. Obtain all bilingual word pairs previously unseen in Bilex and whose S3(i,j)> threshold3=>C3
            and update Bilex;
       c. Compute alignment score=>S4; if (S4> threshold4) return C3 otherwise continue;
4. Update comparable document pairs
       a. Find all pairs of glossed Chinese and English documents which contain parallel sentences
            (anchor sentences) from C2=>C4;
       b. Expand C4 by finding documents similar to each of the document in C4;
       c. C:=C4;
       d. Goto 2;
```

Figure 2. Multi-level bootstrapping algorithm

document position assumptions since such information is not always available.

In addition to assumption 10 on the seed sentence pairs, we propose that while better document matching leads to better parallel sentence extraction, better sentence matching leads to better bilingual lexical extraction, better bilingual lexicon yields better glossing words, which improve the document and sentence match. We can iterate this whole process for incrementally improved results using a multi-level bootstrapping algorithm. Figure 2 outlines the algorithm in more detail. In the following sections 3.1-3.4, we describe the four different steps of our algorithm.

3.1 Extract comparable documents

The aim of this step is to extract the Chinese-English documents pairs that are comparable, and therefore should have similar term distributions.

The documents are word segmented with the Language Data Consortium (LDC) Chinese-English dictionary 2.0. The Chinese document is then glossed using all the dictionary entries. Multiple translations of a Chinese word is disambiguated by looking at the context of the sentences this word appears in (Fung et al., 1999).

Both the glossed Chinese document and the English document are then represented in word vectors, with term weighting. We evaluated different combinations of term weighting of each word in the corpus: term freuency (*tf),* inverse document frequency (*idf*), *tf.idf*, the product of a function of *tf* and *idf.* The ”documents” here are sentences. We find that using *idf* alone gives the best sentence pair rank. This is due to the fact that frequencies of bilingual word pairs are not comparable in a non-parallel, quasi-comparable corpus.

Pair-wise similarities are calculated for all possible Chinese-English document pairs, and bilingual documents with similarities above a certain threshold are considered to be comparable. For quasi-comparable corpora, this document alignment step also serves as topic alignment.

3.2 Extract parallel sentences

In this step, we extract parallel sentences from the matched English and Chinese documents in the previous section. Each sentence is again represented as word vectors. For each extracted document pair, the pair-wise cosine similarities are calculated for all possible Chinese-English sentence pairs. Sentence pairs above a set threshold are considered parallel and extracted from the documents.

We have only used one criterion to determine the parallel-ness of sentences at this stage, namely the number of words in the two sentences that are translations of each other. Further extensions are discussed in the final section of this paper.

3.3 Update bilingual lexicon

Step 3 updates the bilingual lexicon according to the intermediate results of parallel sentence extraction.

The occurrence of unknown words can adversely affect parallel sentence extraction by introducing erroneous word segmentations. This is particularly notorious for Chinese to English translation. For example, ”奥 委 会” (“*Olympic Committee*”) is not found in the bilingual lexicon so the Chinese is segmented into three separate words in the original corpus, each word with an erroneous English gloss. Note that this occurs for unknown words in general, not just transliterated words.

Hence, we need to refine bi-lexicon by learning new word translations from the intermediate output of parallel sentences extraction. In this work, we focus on learning translations for name entities since these are the words most likely missing in our baseline lexicon. The Chinese name entities are extracted with the system described in (Zhai et al 2004). New bilingual word pairs are learned from the extracted sentence pairs based on (Fung and Lo 98) as follows:

1. Extract new Chinese name entities (Zhai et al 2004);
2. For each new Chinese name entity:
   - Extract all sentences that it appears in, from the original Chinese corpus, and build a context word vector;
   - For all English words, collect all sentences it appears in from the original corpus, and build the context vectors;
   - Calculate the similarity between the Chinese word and each of the English word vectors

     $$Sim(\vec{C},\vec{E}) = \frac{\sum_{(C_i E_j)\in A} w(C_i).w(E_j)}{\sqrt{\sum_i w(C_j)\sum_j wE_j}}$$

     where ***A*** is the aligned bilexicon pair between the two word vector.

   - Rank the English candidate according to the similarity score.

Sometimes a Chinese named entity might be translated into a multi-word English collocation. In such a case, we search for and accept the English collocation candidate that does appear in the English documents.

Below are some examples of unknown name entities that have been translated (or transliterated) correctly:

皮诺切特.	Augusto Pinochet (*transliteration*)
奋进号	Space Shuttle Endeavor (*translation*)
奥委会	Olympic Committee (*translation*)
内塔尼亚	Benjamin Netanyahu (*transliteration*)

3.4 Update comparable documents

This step replaces the original corpus by the set of documents that are found to contain at least one pair of parallel sentences. Other documents that are comparable to this set are also included since we believe that even though they were judged to be not similar at the document level, they might still contain one or two parallel sentences. The algorithm then iterates to refine document extraction and parallel sentence extraction. An alignment score is computed in each iteration, which counts, on average, how many known bilingual word pairs actually co-occur in the extracted “parallel” sentences. The alignment score is high when these sentence pairs are really translations of each other.

4 Evaluation

We evaluate our algorithm on a quasi-comparable corpus of TDT3 data, which contains various news stories transcription of radio broadcasting or TV news report from 1998-2000 in English and Chinese Channels.

4.2. Baseline method

The baseline method shares the same preprocessing, document matching and sentence matching with our proposed method. However, it does not iterate to update the comparable document set, the parallel sentence set, or the bilingual lexicon.

Human evaluators then manually check whether the matched sentence pairs are indeed parallel. The precision of the parallel sentences extracted is 43% for the top 2,500 pairs, ranked by sentence similarity scores.

4.3 Multi-level bootstrapping

There are 110,000 Chinese sentences and 290,000 English sentences, which lead to more than 30 *billion* possible sentence pairs. Few of the sentence pairs turn out to be parallel, but many are paraphrasing sentence pairs. For example, in the following extracted sentence pair,

- 洪森 将 成为 柬埔寨 的 唯一 首相 。 (*Hun Sen becomes Cambodia ' s sole prime minister*)
- Under the agreement, Hun Sen becomes Cambodia ' s sole prime minister .

the English sentence has the extra phrase "*under the agreement*".

The precision of parallel sentences extraction is 67% for the top 2,500 pairs using our method, which is 24% higher than the baseline. In addition, we also found that the precision of parallel sentence pair extraction increases steadily over each iteration, until convergence.

For another evaluation, we use the bilingual lexical score as described in Section 2.1 again as a measure of the quality of the extracted bilingual sentence pairs from the parallel corpus, comparable corpus, and quasi-comparable corpus. Word pairs common to all corpora are used in the lexical alignment score. Table 2 shows that the quality of the extracted parallel sentences from the quasi-comparable corpus is similar to those from noisy parallel and comparable corpus, even though both are understandably inferior in terms of parallel-ness when compared to the manually aligned parallel corpus. It is worth noting that the lexical alignment score for the extracted sentence pairs from the quasi-comparable corpus is similar to that for the comparable corpus. This is because we must evaluate different corpora by using word pairs that appear in all corpora. This has eliminated many word pairs some of which are likely to contribute significantly to the alignment score.

Corpus	Alignment method	Bilexicon alignment score
Parallel	manual	3.924949
Comparable	DP on sentence position	1.3685069
Comparable	Absolute sentence position	1.0636631
Quasi-comparable	Multi-level bootstrapping	2.649668
Quasi-comparable	Cosine similarity	1.507132

Table 2: Lexical alignment scores of extracted parallel sentences, based on a common lexicon

Figure 3 shows two pairs of parallel sentences from a parallel corpus and a comparable corpus, showing that the latter are closer to bilingual paraphrases rather than literal translations.

Parallel sentence from parallel corpus:

中国 国家 主席 江泽民 抵达 日本 举行 国事访问 。

Chinese president Jiang_Zemin arrived in Japan today for a landmark state visit.

Parallel sentence from comparable corpus:

这 也是 中国 国家 首脑 首次 访问 日本 。

Mr Jiang is the first Chinese head of state to visit the island country.

Figure 3. Example parallel sentences

5 Conclusion

We explore the usability of different bilingual corpora for the purpose of multilingual natural language processing. We compare and contrast a number of bilingual corpora, ranging from the parallel, to comparable, and to non-parallel corpora. The usability of each type of corpus is then evaluated by a lexical alignment score calculated for the bi-lexicon pair in the aligned bilingual sentence pairs.

We compared different alignment assumptions for mining parallel sentences from these different types of bilingual corpora and proposed new assumptions for quasi-comparable corpora. By postulating additional assumptions on seed parallel sentences of comparable documents, we propose a multi-level bootstrapping algorithm to extract useful material, such as parallel sentences and bilexicon, from *quasi-comparable corpora*. This is a completely unsupervised method. Evaluation results show that our approach achieves 67% accuracy and a 23% improvement from baseline. This shows that the proposed assumptions and algorithm are promising for the final objective. The lexical alignment score for the comparable sentences extracted with our unsupervised method is found to be very close to that of the parallel corpus. This shows that our extraction method is effective.

The main contributions of our work lie in steps 3 and 4 and in the iterative process. Step 3 updates the bilingual lexicon from the intermediate results of parallel sentence extraction. Step 4 replaces the original corpus by the set of documents that are found to contain parallel sentences. The algorithm then iterates to refine document extraction and parallel sentence extraction. An alignment score is computed at each iteration, which counts, on average, how many known bilingual word pairs actually co-occur in the extracted parallel sentences. The alignment score is high when these sentence pairs are really translations of each other. By using the correct alignment assumptions, we have demonstrated that a bootstrapping iterative process is also possible for finding parallel sentences and new word translations from comparable corpus.

6 Acknowledgements

This work is partly supported by grants CERG# HKUST6206/03E and CERG#HKUST6213/02E of the Hong Kong Research Grants Council.

References

Regina Barzilay and Noemie Elhadad, *Sentence Alignment for Monolingual Comparable Corpora*, Proc. of EMNLP, 2003, Sapporo, Japan.

Christopher D. Manning and Hinrich Schűtze. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Robert Dale, Hermann Moisl, and Harold Somers (editors), *Handbook of Natural Language Processing.*

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.",

Pascale Fung and Lo Yuen Yee. *An IR Approach for Translating New Words from Nonparallel, Comparable Texts*. In COLING/ACL 1998

Gale, W A and Kenneth W.Church. *A Program for Aligning Sentences in Bilingual Corpora*. Computatinal Linguistics. vol.19 No.1 March, 1993.

Pascale Fung, Liu, Xiaohu, and Cheung, Chi Shun. *Mixed-language Query Disambiguation*. In Proceedings of ACL '99, Maryland: June 1999

Gregory Grefenstette, editor. *Cross-Language Information Retrieval*. Kluwer Academic Publishers, 1998.

Hiroyuki Kaji, *Word sense acquisition from bilingual comparable corpora*, in Proceedings of the NAACL, 2003, Edmonton, Canada, pp 111-118.

Genichiro Kikui. *Resolving translation ambiguity using non-parallel bilingual corpora*. In Proceedings of ACL99 Workshop on Unsupervised Learning in Natural Language

Dragos Stefan Munteanu, Daniel Marcu, 2002. *Processing Comparable Corpora With Bilingual Suffix Trees*. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002).

Dragos Stefan Munteanu, Daniel Marcu, 2004. *Improved Machine Translation Performace via Parallel Sentence Extraction from Comparable Corpora*. In Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL 2004).

Reinhard Rapp. *Identifying word translations in non-parallel texts*. Proceedings of the 33rd Meeting of the Association for Computational Linguistics. Cambridge, MA, 1995. 320-322

Philip Resnik and Noah A. Smith. *The Web as a Parallel Corpus*. Computational Linguistics 29(3), pp. 349-380, September 2003.

Frank Smadja. *Retrieving collocations from text: Xtract*. In Computational Linguistics, 19(1):143-177,1993

Harold Somers. *Bilingual Parallel Corpora and Language Engineering*. Anglo-Indian workshop "Language Engineering for South-Asian languages" (LESAL), (Mumbai, April 2001).

Jean Veronis (editor). *Parallel Text Processing: Alignment and Use of Translation Corpora.* Dordrecht: Kluwer. ISBN 0-7923-6546-1. Aug 2000.

Dekai Wu. *Alignment*. In Robert Dale, Hermann Moisl, and Harold Somers (editors), *Handbook of Natural Language Processing*. 415-458. New York: Marcel Dekker. ISBN 0-8247-9000-6. Jul 2000.

Bing Zhao, Stephan Vogel. *Processing Comparable Corpora With Bilingual Suffix Trees*. In Proceedings of the EMNLP 2002.

Zhai, Lufeng, Pascale Fung. Richard Schwartz, Marine Carpuat and Dekai Wu. *Using N-best list for Named Entity Recognition from Chinese Speech*. To appear in the Proceedings of the NAACL 2004.