Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM

Pascale Fung and Percy Cheung
Human Language Technology Center,
University of Science & Technology (HKUST),
Clear Water Bay, Hong Kong
{pascale,eepercy}@ee.ust.hk

Abstract

We present a method capable of extracting parallel sentences from far more disparate "very-non-parallel corpora" than previous "comparable corpora" methods, by exploiting bootstrapping on top of IBM Model 4 EM. Step 1 of our method, like previous methods, uses similarity measures to find matching documents in a corpus first, and then extracts parallel sentences as well as new word translations from these documents. But unlike previous methods, we extend this with an iterative bootstrapping framework based on the principle of "*find-one-get-more*", which claims that documents found to contain one pair of parallel sentences must contain others even if the documents are judged to be of low similarity. We re-match documents based on extracted sentence pairs, and refine the mining process iteratively until convergence. This novel "*find-one-get-more*" principle allows us to add more parallel sentences from *dissimilar* documents, to the baseline set. Experimental results show that our proposed method is nearly 50% more effective than the baseline method without iteration. We also show that our method is effective in boosting the performance of the IBM Model 4 EM lexical learner as the latter, though stronger than Model 1 used in previous work, does not perform well on data from very-non-parallel corpus.

1. Introduction

Parallel sentences are important resources for training and improving statistical machine translation and cross-lingual information retrieval systems. Various methods have been previously proposed to extract parallel sentences from multilingual corpora. Some of them are described in detail in (Manning and Schűtze, 1999, Wu, 2001, Veronis 2001). The challenge of these tasks varies by the degree of parallel-ness of the input multilingual documents.

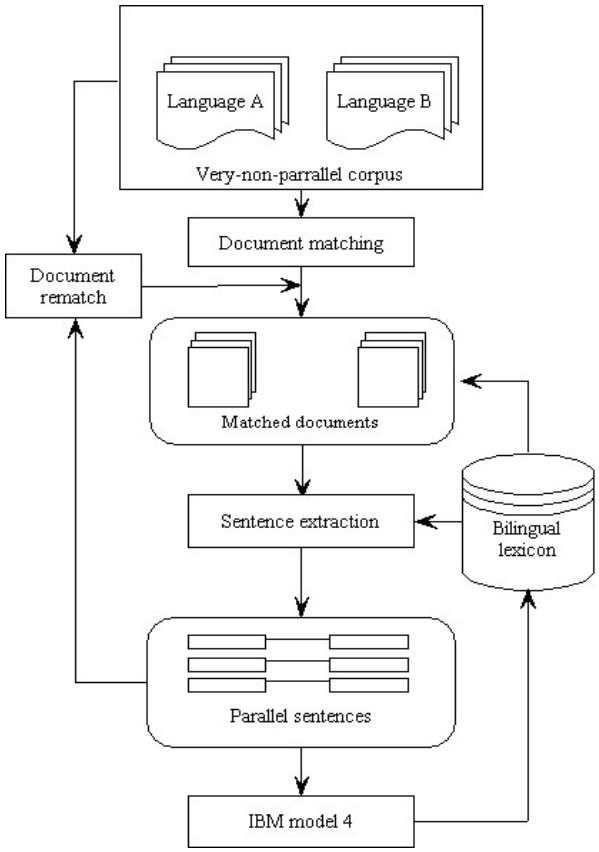

Figure1. Parallel sentence and lexicon extraction via Bootstrapping and EM

The most challenging task is to extract bilingual sentences and lexicon from very-non-parallel data. Recent work (Munteanu et al., 2004, Zhao and Vogel, 2002) on extracting parallel sentences from comparable data, and others on extracting paraphrasing sentences from monolingual corpora (Barzilay and Elhadad 2003) are based on the "find-topic-extract-sentence" principle which claims that parallel sentences only exist in document pairs with high similarity. They all use lexical information (e.g. word overlap, cosine similarity) to match documents first, before extracting sentences from these documents.

However, the non-parallel corpora used so far in the previous work tend to be quite comparable. Zhao and Vogel (2002) used a corpus of Chinese and English versions of news stories from the Xinhua News agency, with *"roughly similar sentence order*

of content". This corpus can be more accurately described as noisy parallel corpus. Barzilay and Elhadad (2003) mined paraphrasing sentences from weather reports. Munteanu et al., (2004) used news articles published within the same 5-day window. All these corpora have documents in the same, matching topics. They can be described as *on-topic* documents. In fact, both Zhao and Vogel (2002) and Barzilay and Elhadad (2003) assume similar sentence orders and applied dynamic programming in their work.

In our work, we try to find parallel sentences from far more disparate, very-non-parallel corpora than in any previous work. Since many more multilingual texts available today contain documents that do not have matching documents in the other language, we propose finding more parallel sentences from off-topic documents, as well as on-topic documents. An example is the TDT corpus, which is an aggregation of multiple news sources from different time periods. We suggest the "find-one-get-more" principle, which claims that as long as two documents are found to contain one pair of parallel sentence, they must contain others as well. Based on this principle, we propose an effective Bootstrapping method to accomplish our task (Figure 1).

We also apply the IBM Model 4 EM lexical learning to find unknown word translations from the extracted parallel sentences from our system. The IBM models are commonly used for word alignment in statistical MT systems. This EM method differs from some previous work, which used a seed-word lexicon to extract new word translations or word senses from comparable corpora (Rapp 1995, Fung & McKeown 1997, Grefenstette 1998, Fung and Lo 1998, Kikui 1999, Kaji 2003).

2. Bilingual Sentence Alignment

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

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 manually aligned sentences, 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 corpus, sometimes also called a "comparable" corpus, contains non-aligned sentences that are nevertheless mostly bilingual translations of the same document. (Fung and McKeown 1997, Kikui 1999, Zhao and Vogel 2002) extracted bilingual word senses, lexicon and parallel sentence pairs from such corpora. A corpus such as Hong Kong News contains documents that 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. Rapp (1995), Grefenstette (1998), Fung and Lo (1998), and Kaji (2003) derived bilingual lexicons or word senses from such corpora. Munteanu et al., (2004) constructed a comparable corpus of Arabic and English news stories by matching the publishing dates of the articles.

Finally, a very-non-parallel corpus is one that contains far more disparate, very-non-parallel bilingual documents that could either be on the same topic (in-topic) or not (off-topic). The TDT3 Corpus is such a 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 around 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 have high similarity. A few of the Chinese and English passages are almost translations of each other. Nevertheless, the existence of a 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. Some of the bilingual sentences are translations of each other, while some others are bilingual paraphrases. Our proposed method is a first approach that can extract bilingual sentence pairs from this type of very-non-parallel corpus.

3. Comparing bilingual corpora

To quantify the parallel-ness or comparability of bilingual corpora, we propose using a lexical matching score computed from the bilingual word pairs occurring in the bilingual sentence pairs. Matching bilingual sentence pairs are extracted from different corpora using existing and the proposed methods.

We then identify bilingual word pairs that appear in the matched sentence pairs by using a bilingual lexicon (bilexicon). The lexical matching score is then defined as the sum of the mutual information score of a known set of 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(Wc,We)* is the co-occurrence frequency of bilexicon pair *(Wc,We)* in the matched sentence pairs. *f(Wc)* and *f(We)* are the occurrence frequencies of Chinese word *Wc* and English word *We*, in the bilingual corpus.

Corpus	Parallel	Comparable	Quasi-Comparable
Lexical matching score	359.1	253.8	160.3

Table 1: Bilingual lexical matching scores of different corpora

Table 1 shows the lexical matching scores of the parallel corpus (Hong Kong Law), a comparable noisy parallel corpus (Hong Kong News), and a very-non-parallel corpus (TDT 3). We can see that the more parallel or comparable the corpus, the higher the overall lexical matching score is.

4. Comparing alignment principles

It is well known that existing work on sentence alignment from parallel corpus makes use of one or multiple of the following principles (Manning and Schűtze, 1999, Somers 2001):

- A bilingual sentence pair are similar in length in the two languages;
- Sentences are assumed to correspond to those roughly at the same position in the other language;
- A pair of bilingual sentences which contain more words that are translations of each other tend to be translations themselves. Conversely, the context sentences of translated word pairs are similar.

For noisy parallel corpora, sentence alignment is based on embedded content words. The word alignment principles used in previous work are as follows:

- Occurrence frequencies of bilingual word pairs are similar;
- The positions of bilingual word pairs are similar;
- Words have one dominant sense/translation per corpus.

Different sentence alignment algorithms based on the above principles can be found in Manning and Schűtze (1999), Somers (2001), Wu (2000), and

1. Initial document matching
For all documents in the comparable corpus D:
Gloss Chinese documents using the bilingual lexicon (Bilex);
For every pair of glossed Chinese document and English documents,
compute *document similarity* =>S(i,j);
Obtain all matched bilingual document pairs whose S(i,j)> threshold1=>D2
2. Sentence matching
For each document pair in D2:
For every pair of glossed Chinese sentence and English sentence,
compute *sentence similarity* =>S2(i,j);
Obtain all matched bilingual sentence pairs whose S2(i,j)> threshold2=>C1
3. EM learning of new word translations
For all bilingual sentences pairs in C1, do:
Compute *translation lexicon probabilities* of all bilingual word pairs =>S3(i,j);
Obtain all bilingual word pairs *previously unseen* in Bilex and whose S3(i,j)> threshold3=>L1, and update Bilex;
Compute *sentence alignment scores*=>S4; if (S4 does not change) return C1 and L1, otherwise continue;
4. Document re-matching
Find all pairs of glossed Chinese and English documents which contain parallel sentences (anchor sentences) from C1=>D3;
Expand D2 by finding documents similar to each of the document in D2;
D2:=D3;
Goto 2;

Figure 2. Bootstrapping with EM

Veronis (2002). These methods have also been applied recently in a sentence alignment shared task at NAACL 2003[1]. We have also learned that as bilingual corpora become less parallel, it is better to rely on lexical information rather than sentence length and position information.

For comparable corpora, the alignment principle made in previous work is as follows:

- Parallel sentences only exist in document pairs with high similarity scores – "find-topic-extract-sentence"

We take a step further and propose a new principle for our task:

- Documents that are found to contain at least one pair of parallel sentences are likely to contain more parallel sentences – "find-one-get-more"

5. Extracting Bilingual Sentences from Very-Non-Parallel Corpora

Existing algorithms such as Zhao and Vogel, (2002), Barzilay and Elhadad, (2003), Munteanu et al., (2004) for extracting parallel or paraphrasing sentences from comparable documents, are based on the "find-topic-extract-sentence" principle which looks for document pairs with high similarities, and then look for parallel sentences only from these documents.

Based on our proposed "find-one-get-more" principle, we suggest that there are other, dissimilar documents that might contain more parallel sentences. We can iterate this whole process for improved results using a Bootstrapping method. Figure 2 outlines the algorithm in more detail. In the following sections 5.1-5.5, we describe the document pre-processing step followed by the four subsequent iterative steps of our algorithm.

5.1. Document preprocessing

The documents are word segmented with the Language Data Consortium (LDC) Chinese-English dictionary 2.0.Then the Chinese document is glossed using all the dictionary entries. When a Chinese word has multiple possible translations in English, it is disambiguated by a method extended from (Fung et al. 1999).

5.2. Initial document matching

This initial step is based on the same "find-topic-extract-sentence" principle as in earlier works. The aim of this step is to roughly match the Chinese-English documents pairs that have the same topic, in order to extract parallel sentences from them. Similar to previous work, comparability is defined by cosine similarity between document vectors.

Both the glossed Chinese document and English are represented in word vectors, with term weights. We evaluated different combinations of term weighting of each word in the corpus: term frequency (tf); inverse document frequency (idf); tf.idf; and 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 probably due to the fact that frequencies of bilingual word pairs are not comparable in a very-non-parallel 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 very-non-parallel corpora, this document-matching step also serves as topic alignment.

5.3. Sentence matching

Again based on the "find-topic-extract-sentence" principle, we extract parallel sentences from the matched English and Chinese documents. Each sentence is again represented as word vectors. For each extracted document pair, 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. Sentence similarity is based on the number of words in the two sentences that are translations of each other. The better our bilingual lexicon is, the more accurate the sentence similarity will be. In the following section, we discuss how to find new word translations.

5.4. EM lexical learning from matched sentence pairs

This step updates the bilingual lexicon according to the intermediate results of parallel sentence extraction. New bilingual word pairs are learned from the extracted sentence pairs based on an EM learning method. We use the GIZA++ (Och and Ney, 2000) implementation of the IBM statistical translation lexicon Model 4 (Brown et al., 1993) for this purpose.

This model is based on the conditional probability of a source word being generated by the target word in the other language, based on EM estimation from aligned sentences. Zhao and Vogel (2002) showed that this model lends itself to adaptation and can provide better vocabulary coverage and better sentence alignment probability estimation. In our

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

work, we use this model on the intermediate results of parallel sentence extraction, i.e. on a set of aligned sentence pairs that may or may not truly correspond to each other.

We found that sentence pairs with high alignment scores are not necessarily more similar than others. This might be due to the fact that EM estimation at each intermediate step is not reliable, since we only have a small amount of aligned sentences that are truly parallel. The EM learner is therefore weak when applied to bilingual sentences from very-non-parallel corpus. We decided to try using parallel corpora to initialize the EM estimation, as in Zhao and Vogel (2002). The results are discussed in Section 6.

5.5. Document re-matching: find-one-get-more

This step augments the earlier matched documents by the "find-one-get-more" principle. From the set of aligned sentence pairs, we look for other documents, judged to be dissimilar in the first step, that contain one or more of these sentence pairs. We further find other documents that are similar to each of the monolingual documents found. This new set of documents is likely to be off-topic, yet contains segments that are on-topic. Following our new alignment principle, we believe that these documents might still contain more parallel sentence candidates for subsequent iterations. The algorithm then iterates to refine document matching and parallel sentence extraction.

5.6. Convergence

The IBM model parameters, including sentence alignment score and word alignment scores, are computed in each iteration. The parameter values eventually stay unchanged and the set of extracted bilingual sentence pairs also converges to a fixed size. The system then stops and gives the last set of bilingual sentence pairs as the final output.

6. Evaluation

We evaluate our algorithm on a very-non-parallel 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. We compare the results of our proposed method against a baseline method that is based on the conventional, "find-topic-extract-sentence" principle only. We investigate the performance of the IBM Model 4 EM lexical learner on data from very-non-parallel corpus, and evaluate how our method can boost its performance. The results are described in the following sub-sections.

6.1. Baseline method

Since previous works were carried out on different corpora, in different language pairs, we cannot directly compare our method against them. However, we implement a baseline method that follows the same "find-topic-extract-sentence" principle as in earlier work. The baseline method shares the same preprocessing, document matching and sentence matching steps 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 manually check whether the matched sentence pairs are indeed parallel. The precision of the parallel sentences extracted is 42.8% for the top 2,500 pairs, ranked by sentence similarity scores.

6.2. Bootstrapping performs much better

There are 110,000 Chinese sentences and 290,000 English sentences in TDT3, which lead to more than 30 billion possible sentence pairs. Few of the sentence pairs turn out to be exact translations of each other, but many are bilingual paraphrases. For example, in the following extracted sentence pair, the English sentence has the extra phrase "under the agreement", which is missing from the Chinese sentence:

- 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.

Another example of translation versus bilingual paraphrases is as follows:

- 中国国家主席江泽民抵达日本举行国事访问
  (*The Chinese president Jiang Zemin arrived in Japan today for a state visit)*
  (Translation) Chinese president Jiang Zemin arrived in Japan today for a landmark state visit.
- 这也是中国国家首脑首次访问日本*(This is a first visit by a Chinese head of state to Japan)*
  (Paraphrase) Mr Jiang is the first Chinese head of state to visit the island country.

The precision of parallel sentences extraction is 65.7% for the top 2,500 pairs using our method, which has a 50% improvement over the baseline. In addition, we also found that the precision of parallel sentence pair extraction increases steadily over each iteration, until convergence.

6.3. Bootstrapping can boost a weak EM lexical learner

In this section, we discuss experimental results that lead to the claim that our proposed method can boost a weak IBM Model 4 EM lexical learner.

6.3.1. EM lexical learning is weak on bilingual sentences from very-non-parallel corpora

We compare the performances of the IBM Model 4 EM lexical learning on parallel data (130k sentence pairs from Hong Kong News) and very-non-parallel data (7200 sentence pairs from TDT3) by looking at a common set of source words and their top-N translation candidates extracted. We found that the IBM Model 4 EM learning performs much worse on TDT3 data. Figure 3 shows that the EM learner performs about 30% worse on average on the TDT3 data.

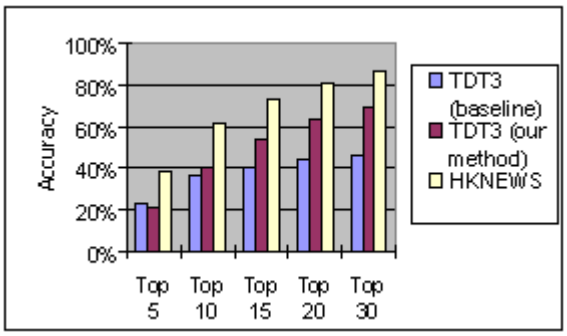

Figure 3. EM lexical learning performance

6.3.2. Multilevel Bootstrapping is significantly better than adaptation data in boosting the weak EM lexical learner

Since the IBM model parameters can be better estimated if the input sentences are more parallel, we have tried to add parallel sentences to the extracted sentence pairs in each iteration step, as proposed by Zhao and Vogel (2002). However, our experiments showed that adding parallel corpus gives no improvement on the final output. This is likely due to (1) the parallel corpus is not in the same domain as the TDT corpus; and (2) there are simply not enough parallel sentences extracted at each step for the reliable estimation of model parameters.

In contrast, Figure 3 shows that when we apply Bootstrapping to the EM lexical learner, the bilingual lexicon extraction accuracy is improved by 20% on the average, evaluated on top-N translation candidates of the same source words, showing that our proposed method can boost a weak EM lexical learner even on data from very-non-parallel corpus.

6.4. Bootstrapping is significantly more useful than new word translations for mining parallel sentences

It is important for us to gauge the effects of the two main ideas in our algorithm, Bootstrapping and EM lexicon learning, on the extraction parallel sentences from very-non-parallel corpora. The baseline experiment shows that without iteration, the performance is at 42.8%. We carried out another set of experiment of using Bootstrapping where the bilingual lexicon is not updated in each iteration. The bilingual sentence extraction accuracy of the top 2500 sentence pairs in this case dropped to 65.2%, with only 1% relative degradation.

Based on the above, we conclude that EM lexical learning has little effect on the overall parallel sentence extraction output. This is probably due to the fact that whereas EM does find new word translations (such as 皮诺切特/Pinochet), this has little effect on the overall glossing of the Chinese document since such new words are rare.

7. Conclusion

Previous work on extracting bilingual or monolingual sentence pairs from comparable corpora has only been applied to documents that are within the same topic, or have very similar publication dates. One principle for previous methods is "find-topic-extract-sentence" which claims that parallel or similar sentences can only be found in document pairs with high similarity. We propose a new, "find-one-get-more" principle which claims that document pairs that contain at least one pair of matched sentences must contain others, even if these document pairs do not have high similarity scores. Based on this, we propose a novel Bootstrapping method that successfully extracts parallel sentences from a far more disparate and very-non-parallel corpus than reported in previous work. This very-non-parallel corpus, TDT3 data, includes documents that are off-topic, i.e. documents with no corresponding topic in the other language. This is a completely unsupervised method. Evaluation results show that our approach achieves 65.7% accuracy and a 50% relative improvement from baseline. This shows that the proposed method is promising. We also find that the IBM Model 4 lexical learner is weak on data from very-non-parallel corpus, and that its performance can be boosted by our Multilevel Bootstrapping method, whereas using parallel corpus for adaptation is not nearly as useful.

In addition, we compare and contrast a number of bilingual corpora, ranging from the parallel, to

comparable, and to very-non-parallel corpora. The parallel-ness of each type of corpus is quantified by a lexical matching score calculated for the bi-lexicon pair distributed in the aligned bilingual sentence pairs. We show that this scores increases as the parallel-ness or comparability of the corpus increases.

Finally, we would like to suggest that Bootstrapping can in the future be used in conjunction with other sentence or word alignment learning methods to provide better mining results. For example, methods for learning a classifier to determine sentence parallel-ness such as that proposed by Munteanu et al., (2004) can be incorporated into our Bootstrapping framework.

References

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

Peter F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. *The mathematics of statistical machine translation: parameter estimation*, in Computational Linguistics, 19-2, 1993.

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

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

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

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

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

Dragos Stefan Munteanu, Alexander Fraser, Daniel Marcu, 2004. *Improved Machine Translation Performance 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).

Franz Josef Och and Hermann Ney. I*mproved statistical alignment models*, in Proceedings of ACL-2000.

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. *Adaptive Parallel Sentences Mining from Web Bilingual News Collections*. In Proceedings of the IEEE Workshop on Data Mining 2002.