Sentence Alignment in Parallel, Comparable, and Quasi-comparable Corpora

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

Abstract

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

We compare different methods of mining parallel sentences and bilingual lexicon from bilingual corpora. These methods make several sentence-level assumptions on the bilingual corpora. We have found that some of them are applicable to bilingual parallel documents but non-applicable to non-parallel, comparable documents. None of the sentence-level assumptions can be made about non-parallel and quasi-comparable corpora. The latter contain bilingual documents that may or may not be on the same topic.

By postulating additional assumptions on comparable documents, we propose a completely unsupervised method to extract useful material, such as parallel sentences and bilexicons, from quasi-comparable corpora. 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.

Introduction

There is an explosively increasing amount of new content being loaded to the Internet every day. These online resources constitute practically an unlimited amount of raw material of corpora for natural language processing, such as multilingual information extraction, question answering, machine translation, and so on (Resnik & Smith, 2003)

One of the most challenging tasks in multilingual information extraction is to identify the comparable documents that are more or less within the same topic. This requires the comparison of documents in different languages that are *not* translations of each other.

What is a comparable document? EAGLES Guidelines1 gives a definition of "comparable corpora".

> "*A comparable corpus is one which selects similar texts in more than one language or variety. There is as yet no agreement on the nature of the similarity, because there are very few examples of comparable corpora.*"

The degree of comparability of different documents varies, but we believe that the more comparable the corpora are, it is more useful for various NLP research task.

We can view both parallel and non-parallel corpora as "extreme" cases of comparable corpora. Our objective is to extract parallel sentences from non-parallel, and quasi-comparable corpora.

In this paper, we describe a method for quantifying the comparability of a bilingual corpus. Then we compare different methods for mining parallel sentences and bilingual lexicon, from bilingual corpora with different degrees of comparability. These methods are based on different assumptions about the characteristics of bilingual corpora. We have found that some assumptions for bilingual parallel documents are non-applicable to non-parallel documents. Finally, by postulating additional assumptions on comparable documents, we propose a completely unsupervised method to extract useful material, such as parallel sentences and bilexicons, from a quasi-comparable corpus

Bilingual Corpora

We compare and contrast bilingual corpora, ranging from the parallel, non-parallel but comparable, and to non-parallel and not very comparable corpora—quasi-comparable corpora.

The Hong Kong Laws Corpus is a parallel corpus with sentence level alignment; it is used as parallel sentence source 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 for the last decade and can be achieved with many off-the-shelf, publicly available alignment tools.

Previous works have extracted bilingual word senses, lexicon and parallel sentence pairs from noisy parallel corpora. This type of corpora is often called comparable corpora. Corpora like the Hong Kong News Corpus, and the Xinhua News Corpus are in fact rough translations of each other, focused on the same thematic topics, with some insertions and deletions of paragraphs. Sentence and bilingual extraction methods from such corpora can be found in (Fung & McKeown, 1995; Fung & Lo, 1998; Zhao & Vogel, 2002).

On the other hand, TDT3 Corpus is a 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 60 different topics. 1,200 Chinese and 4,500 English documents are manually labeled as relevant to a topic and are *in-topic*. The remaining documents are labeled as *off-topic* since they are only weakly relevant to a topic or irrelevant to all topics. The high percentage of off-topic gives rise to more variety of sentences in term of content and structure. From the *in-*

[1] http://www.ilc.cnr.it/EAGLES96/home.html

topic documents, most are found to be comparable. A few of the Chinese and English document are almost parallel document and contain some parallel sentences. Nevertheless, the existence of considerable amount of *off-topic* document makes the whole corpus quasi-comparable. The TDT 3 corpus also contains 110,000 Chinese, 290,000 English sentences, giving more than 30 billion possible sentence pairs. A very small portion of the sentence pairs will turn out to be parallel, but many are sentence pairs describing comparable content, with some addition or deletion of minor information or details. The objective of our proposed method is to automatically identify documents that are on the same topic, and then extract parallel sentence pairs from these documents.

Comparing Bilingual Corpora

We argue that the usability of bilingual corpus is determined by how well the sentences are aligned. We postulate that if the sentence pairs in the corpus are indeed translations of each other, then bilingual word pairs identified in the dictionary will co-occur frequently in this corpus.

Lexical alignment score is defined as the sum of the mutual information score of the bilingual lexicon (bilexicon):

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

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

where *$f(W_c,W_e)$* is the co-occurrence frequency of bilexicon pair *(W_c,W_e)* in the aligned sentence pairs. *$f(W_c)$*, *$f(W_e)$* is the occurrence frequency of Chinese word W_c and English word W_e, in the respective language sentences set.

We use different alignment methods to extract bilingual parallel sentence pairs from the parallel corpus (Hong Kong Law), a comparable noisy parallel corpus (Hong Kong News), and a non-parallel, quasi-comparable corpus (TDT 3). The lexical alignment scores are computed from the extracted sentence pairs and shown in the following table. 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. Corpus comparability

In the following section, we describe the different methods we use for extracting bilingual sentence pairs from parallel, comparable, and not-so-comparable corpora.

Comparing Alignment Methods

All previous work on sentence alignment from parallel corpus makes use of one or multiple of the following assumptions:

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 for bilingual word pairs are made 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 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), and Veronis (2002). These methods have also been applied recently in a sentence alignment shared task at NAACL 2003[2]. 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 work are based soly on bilexical context assumption (Fung & McKeown, 1995; Rapp, 1995; Grefenstette, 1998; Fung & Lo, 1998; Kikui, 1999; Barzilay & Elhadad, 2003; Masao & Hitoshi, 2003; Kenji & 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.

More recent works on mining parallel sentences from non-parallel comparable corpus are (Munteanu & Marcu, 2002; Zhao & Vogel, 2002). Both work use a translation-model based alignment model trained from parallel corpus and adaptively extract more parallel sentences and bilingual lexicon in the comparable corpus. There are several differences between the two methods. Zhao and Vogel (2002) used a generative statistical machine translation alignment model, while Munteanu and Marcu (2002) used suffix trees. In Zhao and Vogel (2002), the comparable corpus consists of Chinese and English versions of new stories from the Xinhua News agency, while Munteanu and Marcu (2002) used unaligned segments from the French-English Hansard corpus and finds parallel sentences among them.

Existing algorithms (Barzilay & Elhadad, 2003; Masao & Hitoshi, 2003; Kenji & Hideki, 2002), for extracting parallel sentences from comparable documents follow similar steps: firstly extract comparable documents and then extract parallel corpus from comparable documents. They 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

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

similarity computation include word overlap count, cosine similarity, and classification scores of a binary classifier trained from parallel corpora, generative alignment classifier.

We propose a method to find parallel sentences and new word translations from unequal number of sentences in news stories in Chinese and English. In our work, we use simple cosine similarity measures and we dispense with using parallel corpora to train an alignment classifier.

An Alignment Method for Quasi-comparable Corpora

In addition to the bi-lexical context assumption described in the previous section, we postulate an additional assumption about non-parallel, quasi-comparable corpus:

- Bi-lexicon translation probability: Bilingual lexicon with better translation probabilities can improve bilingual document (sentence) matching.
- Topic: Documents and passages that are on the same topic tend to contain parallel or comparable sentences;
- 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.

Based on these assumptions, we propose a first method in extracting useful material from quasi-comparable corpora.

Similar to the iterative process in statistical word alignment methods, we propose that while better document matching leads to better parallel sentence extraction, better sentence matching leads to improved bilingual lexical extraction, the latter in turn improves the document and sentence matches. We propose a multi-level bootstrapping algorithm that iteratively improves the quality of the parallel sentences extracted.

Multi-level Bootstrapping

Step 1: Extract Comparable Documents

The aim of this step is to extract the Chinese-English document pairs that are similar in term distributions.

The documents are word segmented with the Language Data Consortium (LDC) Chinese-English dictionary 2.0. Then the Chinese documents are glossed with the same dictionary. When a Chinese word has multiple possible translations, it is disambiguated with a cohesion scores based method (Gao et al., 2001). Both the glossed Chinese document and English are represented in vector forms, in which the inverse document (where a "document" is a single sentence) frequency is used as the term weight.

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.

Step 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.

Step 3: Update the Bilingual Lexicon

The occurrence of unknown words can adversely affect parallel sentence extraction by introducing erroneous word segmentations. Hence, we need to refine the 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 first (Zhai et al., 2004). Translations of these terms are learned from the extracted sentence pairs based on (Fung & Lo, 98) as follows:

Step 4: Refine 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.

Evaluation

We have evaluated our algorithm on a comparable corpus of TDT3 data. We use our method and a baseline method to extract parallel sentences from this corpus and manually examine the precision of these parallel sentences.

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. . The precision of parallel sentence extract is 43% for the top 2,500 ranked pair. For our approach, the precision of extracted parallel sentences is 67% for the top 2,500 ranked pair, which is 24% higher. In addition, we also found that the precision of parallel sentence pair extraction increases steadily over each iteration in our method, until convergence.

The main contribution of the unsupervised multi-level bootstrapping is in steps 3 and 4 and in the iterative process. The iterative lexicon-sentence alignment process has been previously applied to alignment tasks from parallel corpus. 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.

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. A lexical alignment score calculated for the bi-lexicon pair distributed in the aligned bilingual sentence pairs then evaluates the usability of each type of corpus.

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 bilexicons, 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 our 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.

References

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

Pascale Fung and Kathleen Mckeown. (1997). 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. (1998). "An IR Approach for Translating New Words from Nonparallel, Comparable Texts". In Coling 1998

Jianfeng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, Changning Huang. (2001). "Improving Query Translation for Cross-Language Information Retrieval using Statistical Models. In SIGIR'01 September 9-12,2001, New Orleans, Louisiana, USA.

Gregory Grefenstette, editor. (1998). "Cross-Language Information Retrieval". Kluwer Academic Publishers, 1998.

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

Genichiro Kikui. (1999). 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. (1999). Foundations of Statistical Natural Language Processing. The MIT Press.

Kenji Matsumoto and Hideki Tanaka. (2002) Automatic alignment of Japanese and English Newspaper articles using an MT system and a bilingual Company name dictionary. In LREC-2002, pages 480-484

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).

Reinhard Rapp. (1995). 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. (2003) " The Web as a Parallel Corpus", Computational Linguistics 29(3), pp. 349-380, September 2003.

Masao Utiyama and Hitoshi Isahara. (2003). Reliable measures for aligning Japanese-English news articles and sentences. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan.

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

Dekai Wu. (2000). 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. (2002). Processing Comparable Corpora With Bilingual Suffix Trees, In Proceedings of the ICSLP 2002.

Zhai, Lufeng, Pascale Fung, Richard Schwartz, Marine Carpuat and Dekai Wu. (2004). Using N-best list for Named Entity Recognition from Chinese Speec. In the Proceedings of the NAACL 2004 , to appear