An Empirical Study on Web Mining of Parallel Data

Gumwon Hong[1], Chi-Ho Li[2], Ming Zhou[2] and Hae-Chang Rim[1]

[1]Department of Computer Science & Engineering, Korea University
{gwhong,rim}@nlp.korea.ac.kr

[2]Natural Language Computing Group, Microsoft Research Asia
{chl,mingzhou}@microsoft.com

Abstract

This paper[1] presents an empirical approach to mining parallel corpora. Conventional approaches use a readily available collection of comparable, non-parallel corpora to extract parallel sentences. This paper attempts the much more challenging task of directly searching for high-quality sentence pairs from the Web. We tackle the problem by formulating good search query using 'Learning to Rank' and by filtering noisy document pairs using IBM Model 1 alignment. End-to-end evaluation shows that the proposed approach significantly improves the performance of statistical machine translation.

1 Introduction

Bilingual corpora are very valuable resources in NLP. They can be used in statistical machine translation (SMT), cross language information retrieval, and paraphrasing. Thus the acquisition of bilingual corpora has received much attention.

Hansards, or parliamentary proceedings in more than one language, are obvious source of bilingual corpora, yet they are about a particular domain and therefore of limited use. Many researchers then explore the Web. Some approach attempts to locate bilingual text within a web page (Jiang et al., 2009); some others attempt to collect web pages in different languages and decide the parallel relationship between the web pages by means of *structural cues,* like existence of a common ancestor web page, similarity between URLs, and similarity between the HTML structures (Chen and Nie, 2000; Resnik and Smith, 2003; Yang and Li, 2003; Shi et al., 2006). The corpora thus obtained are generally of high quality and wide variety in domain, but the amount is still limited, as web pages that exhibit those structural cues are not abundant.

Some other effort is to mine bilingual corpora by *textual means* only. That is, two pieces of text are decided to be parallel merely from the linguistic perspective, without considering any hint from HTML markup or website structure. These approaches (Zhao and Vogel, 2002; Utiyama and Isahara 2003; Fung and Cheung, 2004; Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2009) share roughly the same framework:

Phase 1: Document Pair Retrieval

1) documents in some target language (TL) are stored in some database;
2) each document in some source language (SL) is represented by some TL keywords;
3) the TL keywords in (2) are used to assign some TL documents to a particular SL document, using some information retrieval (IR) technique. For example, Munteanu and Marcu (2005) apply the Lemur IR toolkit, Utiyama and Isahara (2003) use the BM25 similarity measure, and Fung and Cheung (2004) use cosine similarity. Each TL document pairs up with the SL document to form a candidate parallel document pair.

Phase 2: Sentence Pair Extraction

1) sentence pairs can be obtained by running sentence alignment over all candidate document pairs (or a selection of them) (Zhao and Vogel, 2002; Utiyama and Isahara, 2003);
2) sentence pairs can also be selected, by some classifier or reliability measure, from the candidate sentence pairs enumerated from the candidate document pairs (Munteanu and Marcu, 2005).

Note that the primary interest of these approaches is sentence pairs rather than document

[1] This work has been done while the first author was visiting Microsoft Research Asia.

pairs, partially because document pair retrieval is not accurate, and partially because the ultimate purpose of these corpora is SMT training, which is based on sentence pairs. It is found that most of the sentence pairs thus obtained are not truly parallel; rather they are loose translations of each other or they carry partially similar messages. Such bilingual corpora are thus known as *comparable corpora*, while genuinely mutual translations constitute *parallel corpora*.

Note also that all these comparable corpus mining approaches are tested on *closed document collections* only. For example, Zhao and Vogel (2002), Utiyama and Isahara (2003), and Munteanu and Marcu (2005) all acquire their comparable corpora from a collection of news articles which are either downloaded from the Web or archived by LDC. The search of candidate document pairs in such a closed collection is easy in three ways:

1) all the TL documents come from the same news agency and they are not mixed up with similar documents from other news agencies;
2) all the TL documents are news text and they are not mixed up with text of other domains;
3) in fact, the search in these approaches is made easier by applying tricks like date window.

There is no evidence that these methods apply to corpus mining from an open document collection (e.g. the entire Web) without search constraint. The possibility of open-ended text mining is a crucial problem.

This paper focuses on bilingual corpus mining using only textual means. It attempts to answer two questions:

1) Can comparable corpus mining be applied to an open document collection, i.e., the Web?
2) Can comparable corpus mining be adapted to parallel corpus mining?

We give affirmation to both questions. For the first problem, we modify document pair retrieval so that there is no longer a closed set of TL documents. Instead we search for candidate TL documents for a particular SL document from the Web by means of some Web search engine. For the second problem, in Phase 2 we replace the sentence pair classifier by a document pair filter and a sentence alignment module. Based on end-to-end SMT experiments, we will show that 1) high quality bilingual corpora can be mined from the Web; 2) the very first key to Web-mining of bilingual corpus is the formulation of good TL keywords to represent a SL document; 3) a simple document pair filter using IBM Model 1 probabilities is able to identify parallel corpus out of noisy comparable text; and 4) Web-mined parallel corpus, despite its smaller size, improves SMT much more than Web-mined comparable corpus.

2 Problem Setting

Our ultimate goal is to mine from the Web training data for translation from Chinese (SL) to English (TL). As the first step, about 11,000 Chinese web pages of news articles are crawled from some Chinese News sites. Then the task is to search for the English sentences corresponding to those in the selected SL articles. These selected SL news articles all contain cue phrases like "根据外电报道" (*according to foreign media*), as these cue phrases suggest that the Chinese articles are likely to have English counterparts. Moreover, each selected SL article has at least 500 words (empirically determined) since we assume that it is much easier to formulate reliable keywords from a long document than a short one.

3 Document Pair Retrieval

Conventional approaches to comparable corpus mining usually start with document pair retrieval, which assigns to each SL document a set of candidate TL documents. This step is essentially a preliminary search for candidate sentence pairs for further scrutiny in Phase 2. The target is to find document pairs which may contain many good sentence pairs, rather than to discard document pairs which may not contain good sentence pairs. Therefore, *recall is much more emphasized* than precision.

Document pair retrieval in conventional approaches presumes a closed set of TL documents which some IR system can handle easily. In this paper we override this presumption and attempt a much more challenging retrieval task, viz. to search for TL documents among the Web, using the search engines of Google and Yahoo. Therefore we are subject to a much noisier data domain. The correct TL documents may not be indexed by the search engines at all, and even when the target documents are indexed, it requires a more sophisticated formulation of queries to retrieve them.

In response to these challenges, we propose various kinds of queries (elaborated in the following subsections). Moreover, we merge the TL documents found by each query into a big collection, so as to boost up the recall. In case a query fails to retrieve any document, we iteratively drop a keyword in the query until some documents are found. On the other hand, although the document pairs in question are of news domain, we use the general Google/Yahoo web search engines instead of the specific news search engines, because 1) the news search engines keep only a few web pages for all pages about the same news event, and 2) we leave open possibility for correct TL documents to be found in non-news web pages.

3.1 Simple Queries

There are three baseline formulations of queries:

1) Query of translations of SL TF-IDF-ranked keywords ($Q_{SL\text{-}TFIDF}$). This is the method proposed by Munteanu and Marcu (2005). All the words in a SL document are ranked by TF-IDF and the top-N words are selected. Each keyword is then translated into a few TL words by a statistically learned dictionary. In our experiments the dictionary is learned from NIST SMT training data.
2) Query of TF-IDF-ranked machine translated keywords ($Q_{TL\text{-}TFIDF}$). It is assumed that a machine translation (MT) system is better at handling lexical ambiguity than simple dictionary translation. Thus we propose to first translate the SL document into TL and extract the top-N TF-IDF-ranked words as query. In our experiments the MT system used is hierarchical phrase-based system (Chiang, 2007).[2]
3) Query of named entities (Q_{NE}). Another way to tackle the drawback of $Q_{SL\text{-}TFIDF}$ is to focus on named entities (NEs) only, since NEs often provide strong clue for identifying correspondence between two languages. All NEs in a SL document are ranked by TF-IDF, and the top-N NEs are then translated (word by word) by dictionary. In our experiments we identify SL (Chinese) NEs implicitly found by the word segmentation algorithm stated in Gao et al. (2003), and the dictionaries for translating NEs include the same one used for $Q_{SL\text{-}TFIDF}$, and the LDC Chinese/English NE dictionary. For the NEs not covered by our dictionary, we use Google translation service as a back-up.

A small-scale experiment is run to evaluate the merits of these queries. 300 Chinese news web pages in three different periods (each 100) are collected. For each Chinese text, each query (containing 10 keywords) is constructed and submitted to both Google and Yahoo Search, and top-40 returned English web pages for each search are kept. Note that the Chinese news articles are not part of 11,000 pages in section 2. In fact, they do not only satisfy the requirement of length and cue phrases (described in section 2), but they also have another property that they are translated from some English news articles (henceforth target pages) on the Web. Thus they are ideal data for studying the performance of document pair retrieval.

To test the influence of translation quality in document pair retrieval, we also try 'oracle queries', i.e. queries formulated directly from the target pages:

1) OQ_{TFIDF}. This is the query of the top-N TF-IDF-ranked words from the target page.
2) OQ_{NE}. This is the query of the top-N TF-IDF-ranked NEs from the target web page.

We define recall as the proportion of SL documents whose true target pages are found. The comparison between a retrieved page and the target page is done by Longest Common Subsequence (LCS) ratio, defined as the length of the longest common word sequence of two documents divided by the length of the longer of two documents. The threshold 0.7 is adopted as it is strict enough to distinguish parallel document pairs from non-parallel ones.

Table 1 shows the recalls for various queries. It can be seen from Tests 6 and 7 that the largest recall, 85% (within top 40 search results), is achieved when the word distributions in the target web pages are known. In the real scenario where the true English word distribution is not known, the recalls achieved by the simple queries are very unsatisfactory, as shown by Tests 1 to 3. This clearly shows how challenging Web-based mining of bilingual corpora is. Another challenge can be observed in comparing across

[2] We also try online Google translation service, and the performance was roughly the same.

ID	Query	Remote	Near	Recent
1	$Q_{SL\text{-}TFIDF}$	7	6	8
2	$Q_{TL\text{-}TFIDF}$	16	19	32
3	Q_{NE}	16	21	38
4	union(2,3)	27	31	48
5	union(1,2,3)	28	31	48
6	OQ_{TFIDF}	56	66	82
7	OQ_{NE}	62	68	85
8	$Overlap_{TFIDF}$	52	51	74
9	$Overlap_{NE}$	55	62	83

Table 1: Recall (%age) of simple queries. 'Remote' refers to news documents more than a year ago; 'Near' refers to documents about 3 months ago; 'Recent' refers to documents in the last two weeks.

columns, viz. it is much more difficult to retrieve outdated news document pairs. This implies that bilingual news mining must be incrementally carried out.

Comparing Test 1 to Tests 2 and 3, it is obvious that $Q_{SL\text{-}TFIDF}$ is not very useful in document pair retrieval. This confirms our hypothesis that suitable TL keywords are not likely to be obtained by simple dictionary lookup. While the recalls by $Q_{TL\text{-}TFIDF}$ are similar to those by Q_{NE}, the two queries contribute in different ways. Test 4 simply merges the Web search results in Tests 2 and 3. The significantly higher recalls in Test 4 imply that each of the two queries finds substantially different targets than each other. The comparison of Test 5 to Test 4 further confirms the weakness of $Q_{SL\text{-}TFIDF}$.

The huge gap between the three simple queries and the oracle queries shows that the quality of translation of keywords from SL to TL is a major obstacle. There are two problems in translation quality: 1) the MT system or dictionary *cannot produce any translation* for a SL word (let us refer to such TL keywords as 'Utopian translations'); 2) the MT system or dictionary *produces an incorrect translation* for a SL word. We can do very little for the Utopian translations, as the only solution is simply to use a better MT system or a larger dictionary. On the contrary, it seems that the second problem can somewhat be alleviated, if we have a way to distinguish those terms that are likely to be correct translations from those terms that are not. In other words, it may be worthwhile to reorder candidate TL keywords by our confidence in its translation quality.

Tests 8 and 9 in Table 1 show that this hypothesis is promising. In both tests the TF-IDF-based (Test 8) or the NE-based (Test 9) keywords are selected from only those TL words that appear both in the target page and the machine translated text of the source page. In other words, we ensure that the keywords in the query must be correct translations. The recalls (especially the recalls by NE-based query in Test 9) are very close to the recalls by oracle queries. The conclusion is, even though we cannot produce the Utopian translations, document pair retrieval can be improved to a large extent by removing incorrect translations. Even an imperfect MT system or NE dictionary can help us achieve as good document pair retrieval recall as oracle queries.

In the next subsection we will take this insight into our bilingual data mining system, by selecting keywords which are likely to be correct translation.

3.2 Re-ranked Queries

Machine learning is applied to re-rank keywords for a particular document. The re-ranking of keywords is based on two principles. The first one is, of course, the confidence on the translation quality. The more likely a keyword is a correct translation, the higher this keyword should be ranked. The second principle is the representativeness of document. The more representative of the topic of the document where a keyword comes from, the higher this keyword should be ranked. The design of features should incorporate both principles.

The representativeness of document is manifested in the following features for each keyword per each document:

- *TF*: the term frequency.
- *IDF*: the inverted document frequency.
- *TF-IDF*: the product of *TF* and IDF.
- *Title word*: it indicates whether a keyword appears in the title of the document.
- *Bracketed word*: it indicates whether a word is enclosed in a bracket in the source document.
- *Position of first appearance*: the position where a keyword first appears in a document, normalized by number of words in the document.

- *NE types*: it indicates whether a keyword is a person, organization, location, numerical expression, or non NE.

The confidence on translation quality is manifested in the following features:

- *Translation source*: it indicates whether the keyword (in TL) is produced by MT system, dictionary, or by both.
- *Original word*: it indicates whether the keyword is originally written in *English* in the source document. Note that this feature also manifests the representativeness of a document.
- *Dictionary rank*: if the keyword is a NE produced by dictionary, this feature indicates the rank of the NE keyword among all translation options registered in the dictionary.

It is difficult to definitely classify a TL keyword into good or bad translation in absolute sense, and therefore we take the alternative of ranking TL keywords with respect to the two principles. The learning algorithm used is Ranking SVM (Herbrich et al., 2000; Joachims, 2006), which is a state-of-the-art method of the "Learning to rank" framework.

The training dataset of the keyword re-ranker comprises 1,900 Chinese/English news document pairs crawled from the Web[3]. This set is not part of 11,000 pages in section 2. These document pairs share the same properties as those 300 pairs used in Section 3.1. For each English/target document, we build a set T_{ALL}, which contains all words in the English document, and also a set T_{NE}, which is a subset of T_{ALL} such that all words in T_{NE} are NEs in T_{ALL}. The words in both sets are ranked by TFIDF. On the other hand, for each Chinese/source document, we machine-translate it and then store the translated words into a set S, and we also add the dictionary translations of the source NEs into S. Note that S is composed of both good translations (appearing in the target document) and bad translations (not appearing in the target document).

Then there are two ways to assign labels to the words in S. In the first way of labeling (L_{ALL}), the label ***3*** is assigned to those words in S which are ranked among top 5 in T_{ALL}, label ***2*** to those ranked among top 10 but not top 5 in T_{ALL}, ***1*** to those beyond top 10 but still in T_{ALL}, and ***0*** to those words which do not appear in T_{ALL} at all. The second way of labeling, L_{NE}, is done in similar way with respect to T_{NE}. Collecting all training samples over all document pairs, we can train a model, M_{ALL}, based on labeling L_{ALL}, and another model M_{NE}, based on labeling L_{NE}.

The trained models can then be applied to re-rank the keywords of simple queries. In this case, a set S_{TEST} is constructed from the 300 Chinese documents in similar way of constructing S. We repeat the experiment in Section 3.1 with two new queries:

1) $Q_{RANK\text{-}TFIDF}$: the top N keywords from re-ranking S_{TEST} by M_{ALL};
2) $Q_{RANK\text{-}NE}$: the top N keywords from reranking S_{TEST} by M_{NE}.

Again N is chosen as 10.

ID	Query	Remote	Near	Recent
10	$Q_{RANK\text{-}TFIDF}$	18	20	29
11	$Q_{RANK\text{-}NE}$	35	43	54
12	union(10,11)	39	49	63

Table 2: Recall (%age) of re-ranked queries.

The results shown in Table 2 indicate that, while the re-ranked queries still perform much poorer than oracle queries (Tests 6 and 7 in Table 1), they show great improvement over the simple queries (Tests 1 to 5 in Table 1). The results also show that re-ranked queries based on NEs are more reliable than those based on common words.

4 Sentence pair Extraction

The document pairs obtained by the various queries described in Section 3 are used to produce sentence pairs as SMT training data. There are two different methods of extraction for corpora of different nature.

4.1 For Comparable Corpora

Sentence pair extraction for comparable corpus is the same as that elaborated in Munteanu and Marcu (2005). All possible sentence pairs are enumerated from all candidate document pairs produced in Phase 1. These huge number of candidate sentence pairs are first passed to a coarse sentence pair filter, which discards very unlikely candidates by heuristics like sentence

[3] We also attempt to add more training data for re-ranking but the performance remain the same.

length ratio and percentage of word pairs registered in some dictionary.

The remaining candidates are then given to a Maximum Entropy based classifier (Zhang, 2004), which uses features based on alignment patterns produced by some word alignment model. In our experiment we use the HMM alignment model with the NIST SMT training dataset. The sentence pairs which are assigned as positive by the classifier are collected as the mined comparable corpus.

4.2 For Parallel Corpora

The sentence pairs obtained in Section 4.1 are found to be mostly not genuine mutual translations. Often one of the sentences contains some extra phrase or clause, or even conveys different meaning than the other. It is doubtful if the document pairs from Phase 1 are too noisy to be processed by the sentence pair classifier. An alternative way for sentence pair extraction is to further filter the document pairs and discard any pairs that do not look like parallel.

It is hypothesized that the parallel relationship between two documents can be assimilated by the word alignment between them. The document pair filter produces the Viterbi alignment, with the associated probability, of each document pair based on IBM Model 1 (Brown et al., 1993). The word alignment model (i.e. the statistical dictionary used by IBM Model 1) is trained on the NIST SMT training dataset. The probability of the Viterbi alignment of a document pair is the sole basis on which we decide whether the pair is genuinely parallel. That is, an empirically determined threshold is used to distinguish parallel pairs from non-parallel ones. In our experiment, a very strict threshold is selected so as to boost up the precision at the expense of recall.

There are a few important details that enable the document pair filter succeed in identifying parallel text:

1) Function words and other common words occur frequently and so any pair of common word occupies certain probability mass in an alignment model. These common words enable even non-parallel documents achieve high alignment probability. In fact, it is well known that the correct alignment of common words must take into account positional and/or structural factors, and it is beneficial to a simple alignment model like IBM Model 1 to work on data without common words. Therefore, all words on a comprehensive stopword list must be removed from a document pair before word alignment.
2) The alignment probability must be normalized with respect to sentence length, so that the threshold applies to all documents regardless of document length.

Subjective evaluation on selected samples shows that most of the document pairs kept by the filter are genuinely parallel. Thus the document pairs can be broken down into sentence pairs simply by a sentence alignment method. For the sentence alignment, our experiments use the algorithm in Moore (2002).

5 Experiments

It is a difficult task to evaluate the quality of automatically acquired bilingual corpora. As our ultimate purpose of mining bilingual corpora is to provide more and better training data for SMT, we evaluate the parallel and comparable corpora with respect to improvement in Bleu score (Papineni et al., 2002).

5.1 Experiment Setup

Our experiment starts with the 11,000 Chinese documents as described in Section 2. We use various combinations of queries in document pair retrieval (Section 3). Based on the candidate document pairs, we produce both comparable corpora and parallel corpora using sentence pair extraction (Section 4). The corpora are then given to our SMT systems as training data.

The SMT systems are our implementations of phrase-based SMT (Koehn et al., 2003) and hierarchical phrase-based SMT (Chiang, 2007). The two systems employ a 5-gram language model trained from the Xinhua section of the Gigaword corpus. There are many variations of the bilingual training dataset. The B1 section of the NIST SMT training set is selected as the baseline bilingual dataset; its size is of the same order of magnitude as most of the mined corpora so that the comparison is fair. Each of the mined bilingual corpora is compared to that baseline dataset, and we also evaluate the performance of the combination of each mined bilingual corpus with the baseline set.

Bilingual Training Corpus	Phrase-based SMT (PSMT)		Hierarchical PSMT	
	NIST 2005	NIST 2008	NIST 2005	NIST 2008
B1 (baseline)	33.08	21.66	32.85	21.18
B1+comparable(M&M)	33.51(+0.43)	22.71(+1.05)	32.99(+0.14)	22.11(+0.93)
B1+comparable($Q_{RANK-NE}$)	**34.81(+1.73)**	23.30(+1.64)	34.43(+1.58)	22.85(+1.67)
B1+comparable(all simple)	34.74(+1.66)	**23.48(+1.82)**	34.28(+1.43)	**23.18(+2.00)**
B1+comparable(all ranked)	34.79(+1.71)	**23.48(+1.82)**	34.37(+1.52)	23.06(+1.88)
B1+comparable(all query)	34.74(+1.66)	23.19(+1.53)	**34.46(+1.61)**	23.12(+1.94)
B1+parallel($Q_{RANK-NE}$)	34.75(+1.67)	23.37(+1.71)	34.24(+1.39)	23.45(+2.27)
B1+parallel(all simple)	34.99(+1.91)	23.96(+2.30)	34.94(+2.09)	23.35(+2.17)
B1+parallel(all ranked)	34.76(+1.68)	23.41(+1.75)	34.54(+1.69)	23.59(+2.41)
B1+parallel(all query)	**35.40(+2.32)**	**23.47(+1.81)**	**35.27(+2.42)**	**23.61(+2.43)**

Table 4: Evaluation of translation quality improvement by mined corpora. The figures inside brackets refer to the improvement over baseline. The bold figures indicate the highest Bleu score in each column for comparable corpora and parallel corpora, respectively.

The SMT systems learn translation knowledge (phrase table and rule table) in standard way. The parameters in the underlying log-linear model are trained by Minimum Error Rate Training (Och, 2003) on the development set of NIST 2003 test set. The quality of translation output is evaluated by case-insensitive BLEU4 on NIST 2005 and NIST 2008 test sets[4].

5.2 Experimental result

Table 3 lists the size of various mined parallel and comparable corpora against the baseline B1 bilingual dataset. It is obvious that for a specific type of query in document pair retrieval, the parallel corpus is significantly smaller than the corresponding comparable corpus.

The apparent explanation is that a lot of document pairs are discarded due to the document pair filter. Note that the big difference in size of the two comparable corpora by single queries, i.e., $Q_{RANK-NE}$ and M&M, verifies again that re-ranked queries based on NEs are more reliable in sentence pair extraction.

Queries	SP extraction	#SP	#SL words	#TL words
Baseline: B1 in NIST		68K	1.7M	1.9M
M&M	comparable	43K	1.1M	1.2M
$Q_{RANK-NE}$	comparable	98K	2.7M	2.8M
all simple	comparable	98K	2.6M	2.9M
all ranked	comparable	115K	3.1M	3.3M
all query	comparable	135K	3.6M	4.0M
$Q_{RANK-NE}$	parallel	66K	1.9M	1.8M
all simple	parallel	52K	1.5M	1.4M
all ranked	parallel	73K	2.1M	2.0M
all query	parallel	90K	2.5M	2.4M

Table 3: Statistics on corpus size. SP means sentence pair. ‘all simple’, ‘all ranked’, and ‘all query’ refer to the merge of the retrieval results of all simple queries, all re-ranked queries, and all simple and re-ranked queries, respectively; M&M (after Munteanu and Marcu (2005)) refers to $Q_{SL-TFIDF}$.

Table 4 lists the Bleu scores obtained by *augmenting* the baseline bilingual training set *with* the mined corpora. The most important observation is that, despite their smaller size, parallel corpora lead to no less, and often better, improvement in translation quality than comparable corpora. That is especially true for the case where document pair retrieval is based on all five types of query[5]. The superiority of parallel corpora confirms that, in Phase 2 (sentence pair extraction), quality is more important than quantity and thus the filtering of document pair/sentence pair must not be generous.

On the other hand, sentence pair extraction for parallel corpora generally achieves the best result when all queries are applied in document pair retrieval. It is not sufficient to use the more sophisticated re-ranked queries. That means in Phase 1 quantity is more important and we must seek more ways to retrieve as many document pairs as possible. That also confirms the emphasis on recall in document pair retrieval.

Looking into the performance of comparable corpora, it is observed that the M&M query does not effectively apply to Web mining of comparable corpora but the proposed queries do. Any of the proposed query leads to better result than the conventional method, i.e. M&M. Moreover, it can be seen that all four combinations of proposed queries achieve similar per-

[4] It is checked that there is no sentence in the test sets overlapping with any sentences in the mined corpus.

[5] $Q_{SL-TFIDF}$, $Q_{TL-TFIDF}$, Q_{NE}, $Q_{RANK-TFIDF}$, and $Q_{RANK-NE}$

Bilingual Training Corpus	Phrase-based SMT		Hierarchical PSMT	
	NIST 2005	NIST 2008	NIST 2005	NIST 2008
B1 (baseline)	33.08	21.66	32.85	21.18
comparable(M&M)	20.84(-12.24)	14.33(-7.33)	20.65(-12.20)	13.73(-7.45)
comparable($Q_{\text{RANK-NE}}$)	26.78(-6.30)	18.54(-3.12)	27.10(-5.75)	18.02(-3.16)
comparable(all simple)	26.39(-6.69)	18.52(-3.14)	26.40(-6.45)	18.22(-2.96)
comparable(all ranked)	27.36(-5.72)	18.89(-2.77)	27.40(-5.45)	18.72(-2.46)
comparable(all query)	**27.96(-5.12)**	**19.27(-2.39)**	**27.83(-5.02)**	**19.46(-1.72)**
parallel($Q_{\text{RANK-NE}}$)	26.37(-6.71)	18.70(-2.96)	26.47(-6.38)	18.51(-2.67)
parallel(all simple)	25.65(-7.43)	18.69(-2.97)	25.28(-7.57)	18.55(-2.63)
parallel(all ranked)	26.86(-6.22)	18.94(-2.72)	27.10(-5.75)	18.78(-2.40)
parallel(all query)	**27.58(-5.50)**	**19.73(-1.93)**	**28.10(-4.75)**	**19.52(-1.66)**

Table 5: Evaluation of translation quality by mined corpora.

formance. This illustrates a particular advantage of using a single re-ranked query, viz. $Q_{\text{RANK-NE}}$, because it significantly reduces the retrieval time and downloading space required for document pair retrieval as it is the main bottleneck of whole process.

Table 5 lists the Bleu scores obtained by *replacing* the baseline bilingual training set *with* the mined corpora. It is easy to note that translation quality drops radically by using mined bilingual corpus alone. That is a natural consequence of the noisy nature of Web mined data. We should not be too pessimistic about Web mined data, however. Comparing the Bleu scores for NIST 2005 test set to those for NIST 2008 test set, it can be seen that the reduction of translation quality for the NIST 2008 set is much smaller than that for the NIST 2005 set. It is not difficult to explain the difference. Both the baseline B1 training set and the NIST 2005 comprise news wire (in-domain) text only. Although the acquisition of bilingual data also targets news text, the noisy mined corpus can never compete with the well prepared B1 dataset. On the contrary, the NIST 2008 test set contains a large portion of out-of-domain text, and so the B1 set does not gain any advantage over Web mined corpora. It might be that better and/or larger Web mined corpus achieves the same performance as manually prepared corpus.

Note also that the reduction in Bleu score by each mined corpus is roughly the same as that by each other, while in general parallel corpora are slightly better than comparable corpora.

6 Conclusion and Future Work

In this paper, we tackle the problem of mining parallel sentences directly from the Web as training data for SMT. The proposed method essentially follows the corpus mining framework by pioneer work like Munteanu and Marcu (2005). However, unlike those conventional approaches, which work on closed document collection only, we propose different ways of formulating queries for discovering parallel documents over Web search engines. Using learning to rank algorithm, we re-rank keywords based on representativeness and translation quality. This new type of query significantly outperforms existing query formulation in retrieving document pairs. We also devise a document pair filter based on IBM model 1 for handling the noisy result from document pair retrieval. Experimental results show that the proposed approach achieves substantial improvement in SMT performance.

For mining news text, in future we plan to apply the proposed approach to other language pairs. Also, we will attempt to use meta-information implied in SL document, such as “publishing date” or “news agency name”, as further clue to the document pair retrieval. Such meta-information may likely to increase the precision of retrieval, which is important to the efficiency of the retrieval process.

An important contribution of this work is to show the possibility of mining text other than news domain from the Web, which is another piece of future work. The difficulty of this task should not be undermined, however. Our success in mining news text from the Web depends on the cue phrases available in news articles. These cue phrases more or less indicate the existence of corresponding articles in another language. Therefore, to mine non-news corpus, we should carefully identify and select cue phrases.

References

Abdul-Rauf, Sadaf and Holger Schwenk. 2009. Exploiting Comparable Corpora with TER and TERp. In *Proceedings of ACL-IJCNLP 2009 workshop on Building and Using Comparable Corpora*, pages 46–54.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2): 263-311.

Chen, Jiang and Jian-Yun Nie. 2000. Automatic Construction of Parallel Chinese-English Corpus for Cross-Language Information Retrieval. In *Proceedings of NAACL-ANLP*, pages 21-28.

Chiang, David. 2007. Hierarchical Phrase-based Translation. *Computational Linguistics*, 33(2): 202-228.

Fung, Pascale, and Percy Cheung. 2004. Mining very non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing*, pages 57-63.

Gao, Jianfeng, Mu Li, and Changning Huang. 2003. Improved Source-Channel Models for Chinese Word Segmentation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 272-279.

Herbrich, Ralf, Thore Graepel, and Klaus Obermayer. 2000. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pages 115–132. MIT Press, Cambridge, MA.

Jiang, Long, Shiquan Yang, Ming Zhou, Xiaohua Liu, and Qingsheng Zhu. 2009. Mining Bilingual Data from the Web with Adaptively Learnt Patterns. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing*, pages 870-878.

Joachims, Thorsten. 2006. Training Linear SVMs in Linear Time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217-226.

Koehn, Philipp, Franz Och, and Daniel Marcu. 2003. Statistical Phrase-based Translation. In *Proceedings of conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, pages 48-54.

Moore, Robert. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Proceedings of the 5th conference of the Association for Machine Translation in the Americas*, pages 135–144.

Munteanu, Dragos, and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4): 477-504.

Och, Franz J. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160-167.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311-318.

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

Shi, Lei, Cheng Niu, Ming Zhou, and Jianfeng Gao. 2006. A DOM Tree Alignment Model for Mining Parallel Data from the Web. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 489-496.

Utiyama, Masao, 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*, pages 72-79.

Vogel, Stephan. 2003. Using noisy bilingual data for statistical machine translation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 175-178.

Yang, Christopher C., and Kar Wing Li. 2003. Automatic construction of English/Chinese parallel corpora. *Journal of the American Society for Information Science and Technology*, 54(8):730–742.

Zhang, Le. 2004. Maximum Entropy Modeling Toolkit for Python and C++. http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

Zhao, Bing, and Stephan Vogel. 2002. Adaptive Parallel Sentences Mining from Web Bilingual News Collection. In *Proceedings of IEEE international conference on data mining*, pages 745-750.