Mining Parallel Documents Using Low Bandwidth and High Precision CLIR from the Heterogeneous Web

Simon Shi[1], Pascale Fung[1], Emmanuel Prochasson[2], Chi-kiu Lo[1] and Dekai Wu[1]
[1] Human Language Technology Center
Hong Kong University of Science and Technology (HKUST)
Clear Water Bay, Hong Kong
[2] Digital Butter
69 Jervois Street, Sheung Wan, Hong Kong
{eesys,pascale}@ust.hk, emmanuel@butter.com.hk,
{jackielo,dekai}@cs.ust.hk

Abstract

We propose a content-based approach to mine parallel resources from the entire web using cross lingual information retrieval (CLIR) with search query relevance score (SQRS). Our method improves mining recall by going beyond URL matching to find parallel documents from non-parallel sites. We introduce SQRS to improve the precision of mining. Our method makes use of search engines to query for target document given each source document and therefore does not require downloading target language documents in batch mode, reducing computational cost on the local machines and bandwidth consumption. We obtained a very high mining precision (88%) on the parallel documents by the pure CLIR approach. After extracting parallel sentences from the mined documents and using them to train an SMT system, we found that the SMT performance, with 29.88 BLEU score, is comparable to that obtained with high quality manually translated parallel sentences with 29.54 BLEU score, illustrating the excellent quality of the mined parallel material.

1 Introduction

Parallel resources such as bilingual lexicon and sentence translations are typically obtained from translated parallel documents. The web has now grown into an archive of trillions of URLs, heterogeneous in nature, in the last decade. There is a need to readdress the problem of how to mine parallel documents from the web.

We suggest that parallel documents can be mined with high precision from web sites that are not necessarily parallel to each other.

Parallel resources reside on a diverse range of websites which can be classified into the following categories:

Parallel websites: single website with structurally aligned bilingual pages. Typically they are websites of institutions, governments and commercial companies. (e.g. Financial Times Chinese/English, Wall Street Journal Chinese/English). Structure based methods were previously proposed to mine parallel documents from these websites:

Resnik and Smith (2003) used (1) parent pages containing links to versions of one document in different languages and (2) sibling pages contains link to translation of the current documents. They also rely on the URL and anchor text to spot language specific version of documents.

A structural alignment using DOM tree representation was proposed by Shi et al. (2006) to align parallel documents by using HTML structure. They identify the translational equivalent texts and hyperlinks between two parallel DOM trees to find parallel documents.

However, the web is a heterogeneous collection of documents that extend far beyond bilingual and comparable pages with obvious structural features, such as similar URLs or common titles. Structural features only work for bilingual websites or document pairs that are already linked by editors.

Comparable websites: websites that contain parallel content in different languages without any structural relation between document pairs. Press agencies have independent content management systems and editors for publishing news

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 420–428,
Chiang Mai, Thailand, November 8 – 13, 2011.

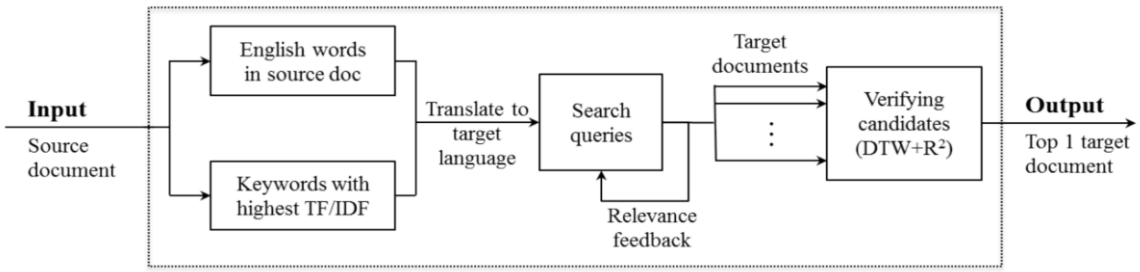

Figure 1. Parallel Document Mining using CLIR with Relevance Feedback

in different languages. (e.g. Reuters China vs. Reuters)

Quasi-comparable websites: independent websites that somewhere contain translated parallel contents. They may contain stories, documentations and books chapters in many languages on different websites. (e.g. Forbes, Fortune)

Instead of structural cues such as URLs, hyperlinks and HTML trees, content based approach are applied to find extra parallel resources from comparable and quasi-comparable websites.

Nie et al. (1999) proposed to download all source language and target language documents and then perform Cross Language Information Retrieval (Grefenstette, 1998) to extract candidate parallel documents. Munteanu and Marcu (2005, 2006) also focused on mining parallel documents from a downloaded collection of news articles, using time stamp alignment and content matching. More recently, Jiang et al. (2009) proposed an adaptive pattern-based bilingual data mining method to mine bilingual web pages for parallel phrases and terms.

Uszkoreit et al. (2010) aligned parallel documents by querying n-gram index built from translation of multilingual documents. All these approaches require a huge local achieve of both source and target documents. This can be very costly when we want to query the entire web.

Moreover, Uszkoreit et al. (2010) makes use of statistical machine translation (SMT) system to translate all documents into target language to build a query index. Due to the complexity of machine translation algorithms, it is still resource wasteful to download all target language documents, machine translate them, then select the desired candidate parallel documents.

Web content is being updated continuously. The above methods need to crawl for all documents in the target language. This is costly in terms of CPU consumption, bandwidth usage and disk storage utilization. This step can be replaced with search engine APIs by several search queries generated from source documents to save CPU and bandwidth consumption.

As most research institutions interested in mining parallel documents do not possess a large number of CPUs or storage on the scale of the world's top search companies, it is also desirable that any site can scale the mining speed and volume according to the computing resources available to them.

To this end, we propose a low bandwidth CLIR method to on the one hand complement structural matching, and on the other hand reduce the complexity of content matching.

Hong et al. (2010) proposed a mining approach on selected Chinese news article containing cue phrases. In non-oracle queries, 45% of the parallel or comparable documents were found among top search results. This is a benchmark in mining precision.

As the parallel resources mined are often times used to improve SMT systems or yield bilingual lexicons, it is desirable that the mining output is of high precision.

2 The Low Bandwidth High Precision Content Based Approach

Our proposed approach (Figure 1) primarily aims to discover parallel documents from all kinds of parallel, comparable or quasi-comparable websites on the World Wide Web. We take advantage of online search engines to find candidate documents thereby saving bandwidth, computational cost and dispenses with crawling for and storing all documents in the target language in an archive.

Content based approach queries the document in target language using keywords from documents in the source language. In our approach, queries are generated from source documents and expanded dynamically by search result quality as feedback. Neither machine translation of the full

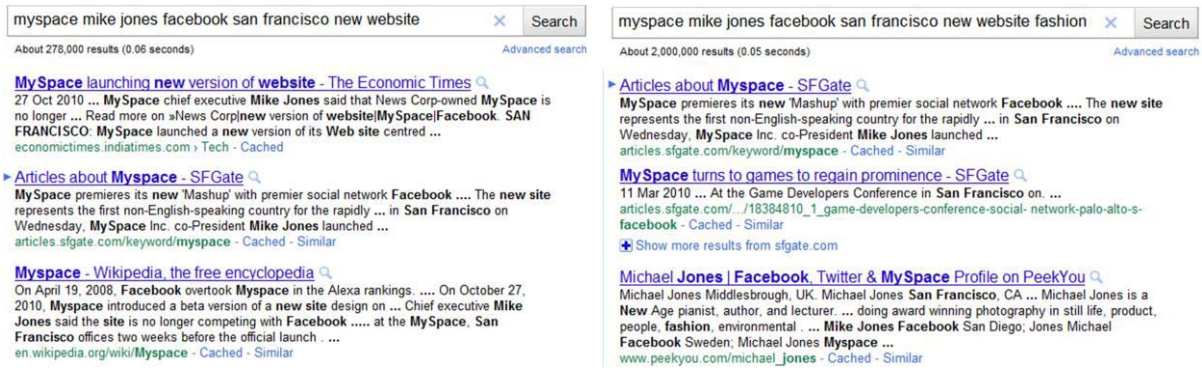

Figure 2. Search Result of Query 1 (Left) and 2 (Right) on Google.com

text no downloading of target documents is needed.

We suggest query expansion feedback score is the key in improving the precision of target documents found. If a source document is found to have no translation in the target language, the system simply returns <not-found>.

2.1 Representing Source Document

We cannot enter documents with thousands of words directly into an online search engine. We need to convert full text into keywords to perform automated queries. A keyword may exist in multiple articles. However, several keywords cam uniquely identify a document if they are grouped together as a keyword set (Jiang et al., 2009).

We then translate each keyword to target language to form the initial query.

There are several reasons why using the translated keyword set as query directly, as proposed by Hong et al. (2010), does not always yields the desired target document:

1) Keyword translation might not correspond to the actual words in the target document;
2) Certain keywords in the target document might have been removed by content editors;
3) There are errors in keyword translation or selection.

It is essential to select appropriate keywords to find the desired target document in a search engine. Two conditions that an appropriate keyword set should satisfy are: (1) they should represent the document exclusively (Jiang et al., 2009) (2) they should have unique or common translation in both languages.

We suggest that words with high TF-IDFs and English words in Chinese text are usually keywords that fulfill both conditions above.

$$K = K_T \bigcup K_E$$

K_T: set of words with high TF-IDF score
K_E: set of English words in Chinese documents

To obtain TF-IDFs that are representative of the keyword in the source document, they are trained from all source documents under the same domain name (e.g. www.ftchinese.com).

Keywords in K_E are more important because most of them are words used in the target document. However, in many cases, there are additional words in K_E so that we cannot find any document by directly searching for K_E. Our method removes keywords with the lowest TF-IDF score from K_E until a non-empty result is obtained.

2.2 Translating Source Documents with Search Query Relevance Score (SQRS)

Search engines use multiple criteria, such as keyword significance, domain popularity, date, popularity, page rank and etc., to return the most relevant documents that match the query. For mining a translated document pair, we need to somehow overcome the impact of page popularity and rank, and aim for content matching only.

Instead of ranking keywords locally and send single query, we take the above search engine criteria into account to amend queries.

To avoid adding erroneously translated keywords and further reduce the amount of undesirable documents downloaded, we introduced the *search query relevance score (SQRS),* defined in Equation 1, that describes how well the search result is and how we can refine the query. The score is determined by comparing the query with highlighted keywords in search result. Generally, a webpage has higher SQRS if the summary contains more keywords that match the query.

Commercial search engines omit some keywords when there is no document in their index

containing all the keywords. In such cases, the rank of documents usually changes significantly.

The following example shows search results of two search queries (Figure 2) generated from the Chinese version of *My Space launching new version of website*[1]. "|" indicates separation of keywords.

Query 1:
myspace | mike jones | facebook | san francisco | new | website

Query 2:
myspace | mike jones | facebook | san francisco | new | website | fashion

In Query 1, the oracle (known) target document was the topmost in search result. The short summary contains every keyword we entered in the query. Rank and SQRSs are shown in Table 1.

SQRS	Search engine omitted kwd	Rank
7.742	-	1*
5.174	web(site)	3
4.951	web(site)	2
4.663	web(site)	4
4.545	web(site)	5

* Target document

Table 1. SQRS of Query 1

In Query 2, we added *fashion* which is the English translation of "新潮" (but the actual English version used *hottest*). The rank of search result changed and each summary omitted at least one keyword in the query (Table 2).

SQRS	Search engine omitted kwd(s)	Rank
6.155	fashion	5*
3.951	web(site) | fashion	1
5.867	website	3
0.871	mike | new | website | fashion	4
-2.921	mike jones | new | website | fashion	2

Table 2. SQRS of Query 2

This phenomenon suggests that the document with all keywords in Query 2 does not exist on the web. The recently added keyword *fashion* must be erroneously translated.

In many similar cases, an erroneously translated keyword can pollute the query quality and decrease the rank of target document. Parallel document mining cannot rely on the document rank of search engine. The system must have a mechanism to detect the problem when expanding the query. Otherwise, a batch of irrelevant documents will be downloaded and need to be filtered out.

We ran experiments to find target documents of 112 randomly selected source documents and compare their SQRSs. 81 or 72.3% target documents have the highest SQRS among other URLs in the search results. It implies the SQRS are an effective measure of query formation and keyword translation.

Source documents	Target documents have largest SQRSs	%
112	81	72.3

Table 3. Result quality and SQRS

Although the query may include multiple translations of a keyword in a bilingual lexicon, the SQRS ensures that there is minimum adverse effect from incorrect translations.

2.3 Query Expansion using SQRS

To improve the precision of the keyword set, we further use SQRS for relevance feedback as shown in Figure 3.

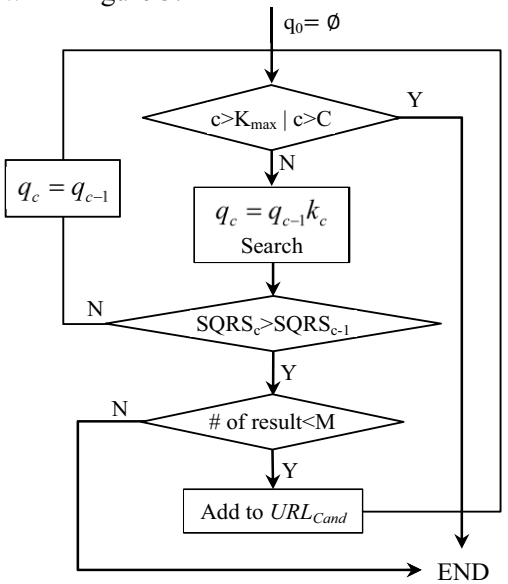

Figure 3. Flowchart of Query Expansion Algorithm

First, we rank the keywords in K_T by their TF-IDF scores. Next, the query is expanded by SQRS. When keyword w is added to current query, we compare the maximum $SQRS_c$ among top n results with the previous highest score $SQRS_p$ without w. w will be discarded from the keywords if $SQRS_c<SQRS_p$ or simply caused an

[1] Source: http://cn.reuters.com/article/CNTechNews/idCNCHINA-3233720101027 on May 10, 2011

$$Q = (k_1 k_2 ... k_C) \quad k_i = (w_{i1} w_{i2} ... w_{iJ}) \quad \Gamma K_i = \{(w_{ia} ... w_{ib}) \mid 0 < a < b \le J\}$$

$$count(c,t) = \#\text{ of occurrence of } c \text{ in } t, \quad \delta(c,t) = \begin{cases} 0 & \text{if } count(c,t) > 0 \\ 1 & \text{if } count(c,t) = 0 \end{cases}$$

$$SQRS(Q,T) = \sum_{c=1}^{C} \left[\sum_{s \in \Gamma K_c} [\log(count(s,T)+1)] - \delta(k_c,T) \right]$$

where Q is the query, k is keyword, w is English word and T is the short text with highlighted keywords in search result.

Equation 1. Definition of SQRS

empty search result. Otherwise, query will be expanded by adding w.

The search engine returns the total number of target documents for each query. If this number is less than a threshold M, we will add the URL of top-ranked documents to the URL_{Cand} list for verification.

To save network bandwidth, the system only considers the top K_{Max} words with the highest TF-IDF scores.

2.4 Document Verification

All candidate document pairs are subjected to a parallelness verification process before output. The system returns <not-found> if a pair failed the verification process. We propose using both dynamic time warping (DTW) and R^2 regression as in (Cheung and Fung, 2004) on every pair of the source and targets document to evaluate their parallelness.

2.4.1 Dynamic Time Warping (DTW) Score

DTW alignment is faster than machine translation (MT). We measure the word level DTW score between source document and target document with local constrain of 5 (Equation 2). Stop words are removed from the English text before DTW processing.

If the there is an entry in the bi-lexicon for a pair of i-th Chinese word and j-th English respectively, the cost of point (i,j) is 0, otherwise 1. The total cost is normalized by maximum number of steps (moves) from (0,0) to (m,n) to convert DTW score to a number between 0 and 1.

Parallel document pairs tend to have a path close to the diagonal line with high DTW score.

$$DTW(i_m, i_n) = c + \min \begin{pmatrix} DTW(i_m, i_n) & \cdots & DTW(i_m - 5, i_n) \\ \vdots & \ddots & \vdots \\ DTW(i_m, i_n - 5) & \cdots & DTW(i_m - 5, i_n - 5) \end{pmatrix}$$

Equation 2. DTW with local distance of 5

Figure 4 shows the DTW paths of a parallel document pair and a non-parallel pair. The parallel documents are aligned and the path with minimum cost is shown along the diagonal of the graph.

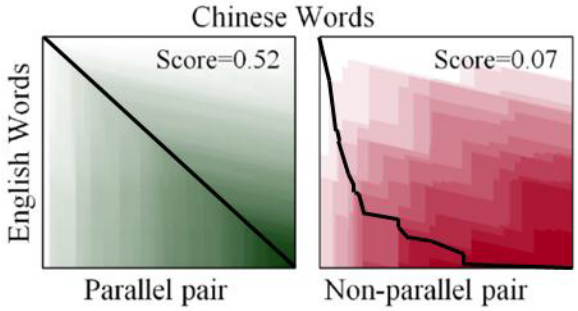

Figure 4. DTW of Parallel and Non-Parallel Pair

Table 4. is the relationship between DTW score and precision of candidate pairs. The precision of output sentences increases if the DTW score threshold is set higher.

DTW	# Pairs	# Parallel	Precision %
>0.45	122	121	99.18
>0.40	224	219	97.77
>0.35	298	288	96.64
>0.30	354	337	95.20
>0.28	389	364	93.57
>0.26	429	389	90.68
>0.25	456	403	88.38
>0.24	488	415	85.04
>0.22	545	417	76.51
>0.20	627	426	67.94

Table 4. DTW and Precision of Candidates Pairs

2.4.2 R^2 Regression

The parallel documents contain parallel sentences that may have different word orders, especially in the case of English and Chinese. The DTW score may be affected by different word order. We propose to use R^2 regression as an additional score to measure the deviation of the matching path of shared words in both documents from the diagonal. (Figure 5)

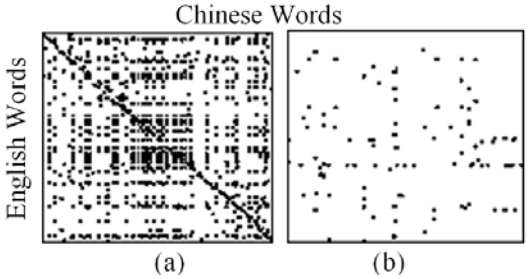

Figure 5. R^2 of Parallel and Non-Parallel Document Pairs

R^2 are normalized by the slope:

$$R^2 = R^2_{score} / Slope$$

2.4.3 Combining DTW and R^2

DTW score helps filter out non-parallel pairs and R^2 is introduced as a supplementary feature to improve the precision of extracted parallel documents.

A comparison of using these measures is shown in Table 5.

	DTW (>0.22)	R^2 (1.0E-5,1)	DTW+R^2
# Pairs	545	534	481
# Parallel	417	403	399
Precision %	76.51	75.47	82.95

Table 5. Mining Precision of DTW and R^2

2.4.4 Structural Features

The final step of verification uses structural features of the document pair candidates:

- **Language:** mined document should be in the target language
- **Absolute size:** mined documents should not have too small/large in file length
- **Size difference:** source and target documents must have similar size
- **Document type:** both documents must be content page in a website

2.5 Find One Get More

Since search engines rank target documents by various criteria, such as the popularity-based page rank, some legitimate bilingual website documents might not be found by our proposed content based method, content based approach using search engines. We propose to supplement our approach with URL matching patterns if the content based method has found several pairs of source and target documents under the same hostname.

Source	# Chinese Docs
ftchinese.com	11,009
cn.wsj.com	3,327
cn.reuters.com	8,570
forbeschina.com	6,281
fortunechina.com	593
Total	**29,780**

Table 6. Source Documents for Pure CLIR Approach

We examine the pairs found by the content based method and look for any parallel pairs coming from the same hostname or whether a pattern can be generalized from these URLs.

We apply this URL pattern to all Chinese pages under this domain.

All pairs found by both methods are subjected to pass the verification process in Section 2.4.

3 Experimental Setup

We evaluate our approach on two sets of experiments.

3.1 Baseline

As a baseline of the content base method, we directly use English words in the original Chinese document as keyword. Then, we add keywords ranked by TF-IDF to query the target document but not perform SQRS to expand query.

Finally, SQRS is used to refine each keyword to get better results.

We use both Google and Bing Search APIs to search the keyword sets. Results from different search engines are merged together by URLs. For each query, we consider eight URLs which is the default number of search engine APIs.

We generalize URL patterns (if any) from document pairs when we find some document pairs by content based method on parallel websites. By *Find One Get More*, we extract more parallel webpages that follow those URL patterns.

3.2 Parallel Document Extraction Accuracy

Source (Chinese) documents in our experiments are news from the following 5 agencies:

Parallel (bilingual) websites:

(1) Financial Times Chinese (ftchinese.com)
(2) Wall Street Journal Chinese (cn.wsj.com)

Parallel website contain both Chinese and English document under the same host and can be aligned with URL matching.

Comparable/quasi-comparable websites:

(3) Reuters China (cn.reuters.com)
(4) Forbes China (forbeschina.com)
(5) Fortune China (fortunechina.com)

Documents on quasi-comparable or comparable websites may have target documents on either the corresponding agencies' global site (e.g. cn.reuters.com and www.reuters.com) or somewhere else. Parallel documents from such websites cannot be found by URL matching.

We applied our content based approach to the above sites to find target documents and evaluate the mining precision.

The percentage of parallel documents that we can successfully find is highly dependent on the type of documents and search engine index. Calculating recall, on the other hand, is only possible for sites we already knew. For comparable or quasi-comparable sites, it is not possible to have the oracle target documents for evaluation because:

1) Some source documents may not have translation in the target language
2) Target language pages may not be indexed by search engines
3) Manual evaluation of all documents for recall calculation is not feasible

In the verification process, we discard the document pairs if:

- DTW score>0.25 (88% precision)
- R^2 score>1.0E-5
- Article size is too small
- Size of source and target too different
- URL is root (/) under hostname
- Text in wrong language

We manually evaluate the effectiveness of our method on randomly selected document pairs. Only parallel document pairs are considered as *correct*.

3.3 Parallel Sentence Extraction

In order to obtain a sentence alignment for pairs of document, we first need to extract the proper content of each page and remove the header and footers that are of little interest and are unlikely to be parallel anyway.

We first segment the documents in sentences and filter out improper ones, such as English sentence containing Chinese characters, or Chinese sentence containing roman characters only. We then use DTW again to find a continuous path in the documents and extract the longest one. The header and footer will generally not align and will be discarded; only the chunk of true alignable content will be preserved.

Using this method, we manage to find the beginning and the end of source and target content and extract it. Then discard pairs of document whose number of extracted sentences are too different. Sentence alignment is performed on the remaining documents using the Champollion ToolKit (Ma, 2006), which is already trained for Chinese-English document pairs.

Finally, we filter all the sentences using a simple word overlap score. Sentences whose lengths are too different or whose word overlap score is too low are discarded, to ensure a high precision at the end.

4 Experimental Results

4.1 Comparison of different methods

	Src doc	Doc pairs	Sent.	Improvement
i	1000	153	2483	Baseline
ii	1000	217	2907	+17.08%
iii	1000	243	3068	+23.56%

i. Direct Search of K_E
ii. Top ranked keywords without SQRS
iii. With SQRS

Table 7. Comparison of different methods

We directly search all English keywords in Chinese documents and found 153 target documents (baseline). Then we search translation of top ranked TF-IDF keywords (ii). With SQRS further improved 23.56% of output sentences comparing to baseline (Table 7). The precision in the three experiments are the same.

4.2 Parallel Document Extraction Accuracy

Among the 29,680 Chinese documents retrieved from the five news agencies, we obtained 7,253 parallel document pairs with 88% precision by content based approach alone.

In many such cases, parallel document pairs are on different websites and be found neither by URL matching nor by content-based methods that use times stamps for matching.

4.3 Find One Get More

With the *Find One Get More* approach, we increase the output of parallel documents from

parallel websites. Table 8 shows that using URL matching can improve the output quantity a lot, compensating for the missing target documents with low page ranks.

Source	# of Doc	CLIR	CLIR+URL
FTChinese	11,009	2,968	9,066
WSJ	3,327	1,002	3,120
Reuters	8,570	1,911	1,911
Forbos	6,281	1,166	1,166
Fortune	593	206	206
Total	**29,780**	**7,253**	**15,469**

Table 8. Output Document Pairs of 4.2 & 4.3

For parallel bilingual websites, the pure content based method can find about 1/3 of the target documents compared to the CLIR+URL method. It shows that, however, our query expansion with relevance feedback approach has higher recall than the 18% produced by the local ranked keywords in Hong et al. (2010).

4.4 Parallel Sentence Extraction for SMT

Among the 15,469 Chinese-English document pairs, we extracted 225,374 parallel sentence pairs with mining precision of over 97% based on human evaluation on randomly selected sentence pairs . We evaluate the quality of those sentences for training machine translation with the Moses SMT engine. We compare the BLEU score obtained with a 4,097,357 sentence pairs corpus, manually aligned (baseline) and the BLEU score obtained with the same corpus, replacing 225,374 sentence pairs by the ones we extracted (CLIR). Results are presented in Table 9, they are evaluated on the NIST MT06 evaluation set.

	BLEU
Baseline	29.54
CLIR	29.88

Table 9. BLEU score obtained for SMT

These results show that our set of sentences, together with a larger parallel corpus, yield results similar to the one obtained with manually aligned sentences only.

The extracted sentences have been processed for rare word translation extraction. (Prochasson and Fung, 2011)

4.5 System Performance and Scalability

We carried out our mining experiments on workstation with 8 states of arts CPU cores. The average time taken for each source document is 30 seconds which is only bottle-necked by the usage limitation of search engine APIs.

As the TF/IDF scores are pre-trained only from the source documents, and our CLIR approach mines target document for each source document individually. Our system can be easily scaled to run in parallel on multiple servers.

5 Conclusion

In this paper, we have proposed a content based CLIR approach to search any part of the Web to find parallel documents without the limitation of URL-matched bilingual web sites. Our method transforms an input source document into a target language query set, then it makes use of search engine APIs, and a proposed query relevance feedback mechanism, and finds the target language document if it exists on the web. We propose a search query relevance score (SQRS) that checks for precision of the query keywords we use to represent the source document. Our proposed method does not require machine translation, nor does it require downloading all documents in the target language into an archive for document matching, thereby saving computational resources.

The query expansion and relevance feedback by SQRS which measures translation correctness ensures high precision in the target document found. Using a verification process, the web documents are further filtered by dynamic time warping and regression scores.

Experimental results show an 88% mining precision on the parallel documents extracted from parallel, comparable and quasi-comparable web sites.

Another experiment on extracting bilingual sentences from the mined documents shows that the sentence extraction adds another layer of verification which further improves the precision from 88% to 97%.

SMT experiments on using our mined parallel sentences, together with a larger baseline training set, to train an SMT system show comparable performances from using our data to that of using manually aligned bilingual sentences. Our system is scalable to run on multiple servers simultaneously and is linear in time to the number of input source documents. It can also be run continuously to discover and mine for newly added web documents that were not there previously. It is also extendable to mine for parallel documents in multiple target languages at the same time.

Acknowledgement

This project is partially funded by a subcontract from BBN, under the DARPA GALE project.

Reference

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 16–23.

Susumu Akamine, Yoshikiyo Kato, Daisuke Kawahara, Keiji Shinzato, Kentaro Inui, Sadao Kurohashi, and Yutaka Kidawara. 2009. Development of a large-scale web crawler and search engine infrastructure. In *Proceedings of the 3rd international Universal Communication Symposium(IUCS'09)*, pages 126–131.

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

Gumwon Hong, Chi-Ho Li, Ming Zhou, and Hae-Chang Rim. 2010. An empirical study on web mining of parallel data. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 474–482, Beijing, China, August. Coling 2010 Organizing Committee.

Rüdiger Gleim, Alexander Mehler, and Matthias Dehmer. 2006. Web corpus mining by instance of wikipedia. In *WAC '06: Proceedings of the 2nd International Workshop on Web as Corpus*, pages 67–74, Morristown, NJ, USA. Association for Computational Linguistics.

Xin Jiang, Yunhua Hu, and Hang Li. 2009. A ranking approach to keyphrase extraction. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'09, pages 756–757, New York, NY, USA. ACM.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proceedings of LREC-2006*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31:477–504, December.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from nonparallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 81–88, Morristown, NJ, USA. Association for Computational Linguistics.

Nie, Michel Simard, Pierre Isabelle, Richard Dur, and Universit De Montral. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81.

Jiang Chen and Jian-Yun Nie. 2000. Parallel web text mining for cross-language information retrieval. In *Recherche d'Informations Assistée par Ordinateur (RIAO)*, pages 62–77.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics,* 29:349–380, September.

Lei Shi, 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*, ACL-44, pages 489–496, Morristown, NJ, USA. Association for Computational Linguistics.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large Scale Parallel Document Mining for Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1101–1109, Beijing, China, August. Coling 2010 Organizing Committee.

Cheung Chi Shun and Pascale Fung. 2004. Unsupervised Learning of a Spontaneous and Colloquial Speech Lexicon in Chinese. In *International Journal of Speech Technology*, Vol. 7, No. 2, pp 173-178, Apr 2004.

Marine Carpuat, Pascale Fung, and Grace Ngai. 2006. Aligning word senses using bilingual corpora. *ACM Transactions on Asian Language and Information Processing*, 5(2):89–120.

Pascale Fung, Emmanuel Prochasson, and Simon Shi. 2010. Trillions of comparable documents, In *LREC Workshop on Building and Using Comparable Corpora*, Malta, May 2010.

Emmanuel Prochasson and Pascale Fung. 2011. Rare word translation extraction from aligned comparable documents. *The 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, Portland, USA.