Automatic Collection of the Parallel Corpus with Little Prior Knowledge

Shutian Ma[1], Chengzhi Zhang[1, 2,*]

1. Department of Information Management, Nanjing University of Science and Technology, Nanjing, China 210094
2. Jiangsu Key Laboratory of Data Engineering and Knowledge Service (Nanjing University), Nanjing, China 210093
{mashutian0608@hotmail.com; zhangcz@njust.edu.cn}

Abstract. As an important resource for machine translation and cross-language information retrieval, collecting large-scale parallel corpus has been paid wide attention. With the development of the Internet, researchers begin to mine the parallel corpora from the multilingual websites. They use some prior knowledge like ad hoc heuristics or calculate the similarity of the webpages structure and content to find the bilingual webpages. This paper presents a method that uses the search engine and little prior knowledge about the URL patterns to get the bilingual websites from the Internet. The method is fast for its low time cost and there is no need for large-scale computation on URL pattern matching. We have collected 88 915 candidate parallel Chinese-English webpages, which average accuracy is around 90.8%. During the evaluation, the true bilingual websites that we found have high similar html structure and good quality translations.

Keywords: Parallel pages, Bilingual website, URL pattern, Web mining

1 Introduction

As an important resource for many natural language processing applications, such as cross-language retrieval [1], machine translation [2], parallel corpus have been one of the key resources. Currently, various websites are bilingual or multilingual. For some international organizations or educational institutions, they will build the webpages in several languages on the site to share the information oversea. The websites in some countries where people speak two or more languages will also have bilingual or multilingual webpages. It becomes more and more common to find the parallel corpora from some known bilingual websites. For example, in Hong Kong governmental websites, there are lots of English-Chinese bilingual webpages, while in Canadian governmental websites, we can find many English-French bilingual webpages as well. Many available systems can find these bilingual websites and they detect the parallel URLs with the URL naming rules, e.g., STRAND [3,4,5], PTMiner [6], PupSniffer [7].

* Corresponding author.

A typical strategy of collecting parallel corpora from bilingual websites involves four fundamental steps: (1) Locating the candidate bilingual websites; (2) Crawling for URLs of candidate parallel web pages; (3) Matching and filtering parallel web pages; (4) Extracting parallel text pairs from the obtained webpages.

Prior knowledge has been widely used when locating the candidate bilingual websites. Most of the systems use the word lists of one language or some anchor information to search on the search engine [5, 6]. Then the search engines will return the URLs of candidate bilingual websites. After crawling for the URLs of webpages, researchers need to find the true parallel webpages within candidate bilingual websites. Due to some constructing rules, when building a bilingual website, some given pairing patterns will be inserted into the parallel webpage URLs, such as the 'english' and 'chinese' in the following pair of bilingual webpages:

Webpage in English: http://www.swd.gov.hk/vs/english/police.html
Webpage in Chinese: http://www.swd.gov.hk/vs/chinese/police.html

In the third step, researchers will use some pre-defined patterns to figure out the correct pairing patterns in candidate bilingual websites. The algorithm of patterns matching results in large-scale computation. Meanwhile the URL pattern-based mining may raise concerns on high bandwidth cost and slow download speed. Some researchers try to find more bilingual webpages via link analysis and they also find out the list of bilingual URL pattern pairs with high credibility [7]. Based on this work, we utilize the search rules of search engine websites and URL patterns with high credibility to obtain bilingual websites from the Internet. The method avoids getting too much irrelevant websites from the search engine and costs less time.

The paper is structured as follows: Section 2 gives a brief review of the related work in the field. Section 3 provides the methodology of our experiment. Section 4 presents the evaluation and analysis that we obtain from the data. Section 5 draws some conclusions and we point out the shortages of our experiment and future works.

2 Related Works

For many data-driven task of NLP, how to get parallel corpora in an efficient way has been the focus in many research projects for years. There are numerous systems automatically acquiring parallel corpora from multilingual websites, for example, STRAND [3,4,5], PTMiner [6]. Many researchers are trying to improve the acquisition performance, such as PupSniffer [7], BITS [8], WPDE [9], the DOM tree alignment model [10] and Bitextor [11]. Some researchers use search engines to find parallel webpages. Microblog has also been one of the resources to get parallel corpora [12]. Among the relevant research, we can find two main types of detecting parallel webpages: the way based on the URL patterns and the way based on the HTML structure. Researchers make use of anchor texts or HTML files in order to find some apparent patterns in the websites or in the URLs which represent different languages, especially the language pattern pairs in the URLs to find more parallel websites. In this paper, we find the parallel webpages via high credible URL patterns using the search engine.

The way based on URL patterns aims at using naming rules of URLs to detect bilingual websites. In early times, researchers are using the re-defined substrings [5, 9] and then comes some automatic ways [13]. Finally, extracted URL pairs are verified based on automatic string pattern recognition instead of prior knowledge [14]. Ye, S.et al [13] made a research of relationship between the content and structure of bilingual websites URLs. The URL patterns and HTML structure have also been combined to find parallel websites [15]. They use the HTML structure to go through the directed graph of bilingual websites simultaneously.

The way based on HTML structure advocates using the structure information of HTML. If the two webpages are parallel to each other, they may have corresponding websites links in the HTML content that connect to another pair of parallel webpages. However, in this way we may just find a little amount of parallel webpages and calculations will be much more than the way based on URL patterns. What's more, in the same bilingual website, content structure of the two languages pages may not be totally same like each other.

In our approach, we search for bilingual websites with language-specific URL substrings and replace the bilingual URL patterns to find a likely candidate pair of parallel URLs. Here we use patterns of high credibility. So what is the credibility of the patterns that we find out in the research? In the research of Kit and Ng [14], they define the linking power of pattern based on the number of URL pairs that it can match. Enhanced algorithms are proposed based on their research to match more bilingual webpages. Zhang and Yao [7] get the global credibility of pattern based on statistical analysis about the link relationship of seed websites available. They also defined the bilingual credibility of a website via link analysis. Depending on the data that we get from their research, we don't need to download all the parallel websites or do any complex pattern matching algorithm within a website.

3 Methodology

This section firstly introduces the idea of exploring parallel websites via the search engine and then it continues to show the detailed steps of the approach.

3.1 Can we collect bilingual websites with little priori knowledge?

There are large-scale of parallel webpages on the Internet. Some researchers explore bilingual websites based on the bilingual URL patterns and they also give the credibility of identified URL patterns [7]. We first make simple search queries with high credible URL patterns, like '*en*', '*eng*', '*english*'. The search query is made according to the search rule of '*inurl:*' in order to find URLs that contains the character, '*en*', '*eng*', '*english*' respectively. We give an example of search query: '*inurl:/en/*'. In this study, we use Google as the search engine. Table 1 shows the numbers of websites after we eliminated the duplicated ones.

Table 1. Numbers of the websites that we get from the search engine

inurl	Websites	Multilingual websites	Multilingual websites with Chinese language	Multilingual websites/All websites	Multilingual with Chinese language/All websites
en	412	323	117	78.40%	28.40%
eng	492	311	106	63.21%	21.54%
english	456	207	93	45.39%	20.39%

From the Table 1, we can see that a certain number of multilingual websites are found by limiting the URL. Multilingual websites containing the Chinese language are around 20% of all the websites from search engine. In the front part of returning results, websites are basically the English pages of those multilingual websites.

3.2 Framework and key technologies

Based on above observations, we can make search rules of 'inurl: source language pattern' to locate webpages in the source language of bilingual websites. Then, we use identified URL pattern pairs to find the existing URL in the target language of bilingual websites. Finally, we will get parallel webpages in a bilingual website. The whole procedure of this method can be divided into two steps:

I. Get the URLs of candidate bilingual websites in the source language;

II. Get the URLs of candidate parallel websites in the target language.

Figure 1 shows the overall framework of our approach.

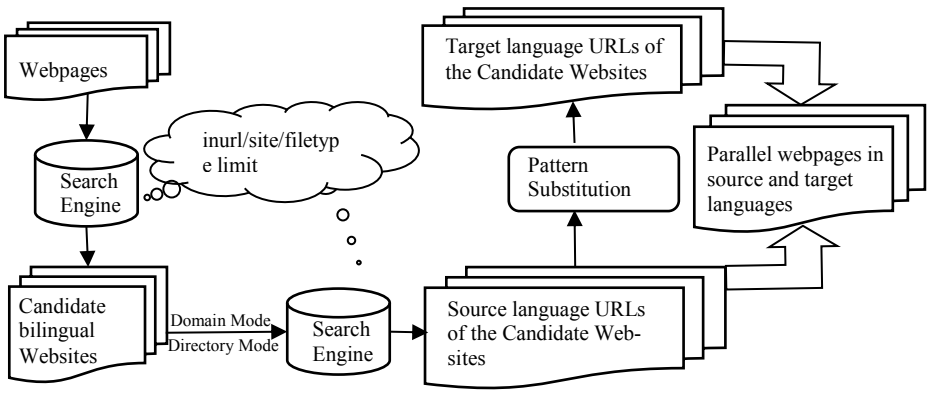

Fig. 1. Framework of our approach

Firstly, with search engine and search rules, we get some candidate bilingual websites, most of these sites are the source language pages of parallel sites.

Secondly, URLs of seed websites were processed in two modes in order to make new search terms. Under the conditions of new search rules, we use the new search terms for a second batch of search and collect the results. From our point of view, the results are actually the websites in the source language of candidate parallel websites.

Thirdly, URLs of the results will have a small substitution. We replaced the source language pattern of each URL with target language pattern to generate a new URL. These pattern pairs are given in the previous research (Zhang and Yao, 2013). If the generated URL is judged to be an existing website, we assume that the new URL and former URL is a pair of parallel websites in the candidate parallel websites. We made a final evaluation for all the URL pairs at last.

Source Language URLs of the Candidate Parallel Websites Detection

One feature of our way is we make use of search rules to find high quality parallel websites from the search engine. When we use the search rules of '*inurl:*', we can find a certain number of multilingual websites from the search engine and some of them have parallel webpages in the site. In this way, we try to add more search rules to define the URL. Then we can get source language URLs of the candidate parallel websites. The main steps of this part are as follows:

i. Search candidate bilingual websites

ii. Search webpages of candidate bilingual websites in the source language

The first step is finding candidate bilingual websites. In this study, the source language is English and target language is Chinese. We aim at the governmental (gov.hk), educational (edu.hk) and organizational (org.hk) types of websites in Hong Kong. To limit the websites types, we add the search rule of '*site:*'. In order to avoid getting too many PDF files, the search rule of '*filetype:html*' is included. Here is an example of the first search query we put in the search box of Google engine.

inurl:en site:gov.hk filetype:html

The second step of this part is finding URLs of candidate parallel websites that in English language. The first thing is reprocessing the URLs of websites that we get from the first search. Then we search again on the search engine with the reprocessed ones. The URLs from the first search are processed in two ways, we call them 'Domain Mode' and 'Directory Mode' here.

(1) Domain Mode: We eliminate the duplicated webpages of seed websites, and keep the second-level domain of each URL. For example, the original URL is'*http://www.immd.gov.hk/en/services/hk-visas/visit-transit.html*'. After processing, the domain mode URL is '*www.immd.gov.hk*'. Then we eliminate the duplicated ones again and check the second-level domain one by one to see if it is subsistent on the Internet. Because of expirations, web server changes or other reasons, some domains are invalid. So we just make use of the accessible ones and a new search query is shown here:

inurl:en site:www.immd.gov.hk filetype:html

In this mode, the file type is defined as html, htm, asp, aspx and php. If the number of the search results is below 100, we get rid of the file type definition. Here is an example of the search query.

inurl:en site:www.immd.gov.hk

(2) Directory Mode: We keep each seed websites URL of the web directory of '*/en/*', '*/eng/*' ,'*/english/*'. We will also give an example here, the original URL is

'*http://www.immd.gov.hk/en/services/hk-visas/visit-transit.html*'. After processing, the directory mode URL is '*www.immd.gov.hk/en/*'. Also we eliminate the duplicated ones again and a new search query is shown here:

site: www.immd.gov.hk/en/ filetype:html

The file type is also defined as html, htm, asp, aspx and php and we collect all the search results during Directory Mode no matter if the number of URLs is below 100. We also do the search without file type definition. You can see that we are trying to find the websites under the web directory of '*www.immd.gov.hk/en/*' as many as possible.

Finally, the search results that we get from the search engine in these two modes are seen as the URLs of candidate websites that in English language. We will use these URLs in the succeeding part.

Target language URLs Generation

Our study is conducted on the basis of identified bilingual URL pairing patterns as described in Zhang and Yao [7]. Their research is conducted on top of the re-implementation of the intelligent web agent to automatically identify bilingual URL pairing patterns as described in Kit and Ng [13]. For some engineering purposes, the web builders will put some static pairing patterns in the pairs of parallel webpages within the same domain. So in the research, they first detected the candidate pairing patterns from candidate URL pairs. For example, the candidate pattern <en,tc> can be detected from the following URL pairs:

Webpage in English: http://www.hkgb.gov.hk/en/news/press_20120227.html
Webpage in Chinese: http://www.hkgb.gov.hk/tc/news/press_20120227.html

Then they use the detected URL patterns to match URLs in a web domain for identifying bilingual webpages [7]. The noisy patterns would be filtered out by thresholding the credibility of a pattern, which can be defined as:

$$C(p,w) = \frac{N(p,w)}{|w|} \tag{1}$$

Where *N(p,w)* is the number of webpages matched into pairs by pattern *p* within website *w*, and |*w*| the size of w in number of webpages.

There are some patterns generalizing across domains. They set the *global credibility* of such a pattern p like this:

$$C(p) = \sum C(p,w)N(p,w) \tag{2}$$

We make use of the patterns with high credibility which are given in their research. The search results that we get from the previous part are the URLs including the character strings like '*en*', '*eng*', '*english*'. For each pattern '*en*', '*eng*', '*english*', we choose 5 candidate pairing patterns respectively to do the substitution according to their credibility in the research of Zhang and Yao (2013). Then we replace the character string like '*en*', '*eng*', '*english*' in the URL with corresponding character string. For instance, the chosen pairing pattern is <en,tc>, <en,b5>,<en,utf-8>…(credibility from high to low order), and we have the URL:

http://www.immd.gov.hk/en/home.html

And a new URL after replacing will be like:

http://www.immd.gov.hk/tc/home.html

Special attention should be paid here. We didn't just replace the '*en*' with '*tc*'. We replace the '*/en/*' with '*/tc/*'. If not it will appear such circumstance, the original URL is:

http://www.csb.gov.hk/mobile/english/info/2047.html

The new URL will be like:

http://www.csb.gov.hk/mobile/tcglish/info/2047.html

So the replacement doesn't make any sense.

Also, for a pairing pattern <english,chinese>, if the '*english*' character string in the URL is in the form of '*/~english/*', we will replace it with '*/~chinese/*'. There are many similar situations during the replacement.

Then new URL is checked after replacing to see whether it exists or not. If the returned http response code of the URL is 200, we assume that the original URL and generated URL is a pair of bilingual parallel websites. Otherwise we keep replacing with the lower credibility bilingual URL pairing patterns until we find an existing website. If the five generated URLs are all checked nonexistent, then we will filter this one and move to the next URL. After the replacements of all the URLs which we get from Section 3.2, we get pairs of URLs which are all existed. One is the URL of candidate websites that in English language. Another one is the URL that we generated with candidate pairing patterns. We assume that these pairs of URL are actually the pairs of parallel webpages in candidate bilingual websites.

4 Experiment and Results Evaluation

4.1 Experimental Data

In this paper, we focused on the governmental, educational and institutional types of sites in Hong Kong when obtaining parallel pages. These three types of the websites are much standard than other types of websites in the content organization and HTML structure. Table 2 shows the number of websites that we get at first.

Table 2. Numbers of the candidate bilingual websites that we get

	site:gov.hk	site:edu.hk	site:org.hk	total
inurl:en	749	1 000	1 000	2 749
inurl:eng	806	992	952	2 750
inurl:english	801	803	671	2 275
Total	2 356	2 795	2 623	7 774

We deal with this websites in two modes: 'Domain Mode' and 'Directory Mode' to get candidate bilingual websites and do the next search.

In the Domain Mode, we keep the second-level domain of each URL from seed websites and eliminate the duplicated ones. Table 3 shows the data of websites that we get after processing. The percentage of English-Chinese websites is higher than the first observation data. Then we just choose the domain that in English-Chinese language to be candidate websites.

In the Directory Mode, we keep each seed websites URL to the web directory of '/*en*/', '/*eng*/' ,'/*english*/', after eliminating the duplicated ones, we get the data in Table 4. We use all the websites in the directory mode to be candidate websites.

Table 3. Numbers of the websites in domain mode

Inurl	Total domains	Total existing domains	English-Chinese domains	English-Chinese / Total domains
en	214	203	193	90.19%
eng	277	264	211	76.17%
english	260	253	139	53.46%

Table 4. Numbers of the websites in directory mode

Web directory	Total Number
www.*./en/	274
www.*./eng/	364
www.*./english/	225

Note: '*' denotes the other strings of the web directory before the /en/,/eng/,/english/

After reprocessing the websites that we get at first, we get candidate websites. Then we make new search queries with search rules and candidate websites limit. After eliminating the duplicated ones in the search results, what we get finally is actually the websites in English language of candidate parallel websites.

The next step is pattern replacing. Table 5 shows the top 5 identified bilingual URL pairing patterns list that we get from the Pupsniffer Evaluation Website[1], a website designed by Zhang and Yao (2013). They released their research data on this website and do the evaluation.

Table 5. Identified bilingual URL pairing patterns list

	en	Credibility	eng	Credibility	english	Credibility
1	en<->tc	13 997.36	eng<->tc	12 869.56	english<->tc_chi	11 436.12
2	en<->b5	5 019.14	eng<->chi	7 824.86	english<->chinese	11 032.46
3	en<->utf-8	4 505.10	eng<->tch	5 281.43	english/<->	261.41
4	en<->ch	3 658.65	eng/<->	1 663.40	english<->traditional	227.37
5	en<->zh	3 460.32	eng<->cht	1 390.22	english<->chi	180.18

Finally we get the pairs of candidate bilingual web pages. The web pages in English language are the search results of candidate websites generated via two modes. The websites in Chinese language are the existing websites after replacing URL patterns. Table 6 shows the number of search results in two modes and the number of existing websites after we checked those replaced URLs. The duplicated ones have

[1] Pupsniffer Evaluation Website, http://mega.ctl.cityu.edu.hk/~czhang22/pupsniffer-eval/

been eliminated here. We can see that more than half of the candidate websites have corresponding websites after URL pattern substitution.

Table 6. Numbers of the websites that we get through the two modes

	Total Websites	Existing Websites	Existing /Total
Domain mode	92 050	55 603	60.41%
Directory mode	109 462	62 034	56.67%
Total	153 683	88 915	57.86%

4.2 Results Evaluation

A web interface was implemented in the Pupsniffer Evaluation Website2 for evaluating the candidate English-Chinese webpage pairs which we finally get. The quality of bilingual webpages found by us is evaluated manually. Two people (one PhD and one master student) took part in this evaluation. Due to the large amount of retrieved pairs, we only randomly sampled and evaluated part of all pairs.

(1) Result of Bilingual Web Pages Collecting

88 915 web pages pairs[3] are found totally via two modes of searching ways. We made an evaluation of 4 460 pairs randomly and the number of the false bilingual web pages pairs is 409. Table 7 shows the precisions by different methods.

Table 7. Performance of different methods

Type of Algorithm	All Pairs	Sampling Ratio	Random Sampling		
			True Pairs	False Pairs	Precision
Kit and Ng(2007)	290 247	3.50%	9 541	603	94.06%
Zhang and Yao (2013)	348 058	4.95%	16 313	910	94.72%
Our method	88 915	5.02%	4 051	409	90.83%

The precision of our method is 90.8%4 and lower than the results of other two methods. However, there are several advantages we need to mention here. Our proposed method has lower time cost. In the other methods, they have to match a mass of the URL patterns within a website in order to find the true parallel webpages. We just make use of the few certain URL patterns pairs based on the previous works. By using the URL patterns of corresponding languages, we don't need to detect the language of websites in advance. Moreover, the method is independent of languages. There is no complex algorithm and we don't need to waste too much time on calculating. All we need is the search engine and a little prior knowledge about URL pairing patterns with high credibility. That is to say, our method is fast and easy while the precision is not low as well.

[2] The login webpage of the Pupsniffer Evaluation Website, http://mega.ctl.cityu.edu.hk/~czhang22/pupsniffer-eval/login.html

[3] The 88 915 web pages pairs result on the Pupsniffer Evaluation Website, http://mega.ctl.cityu.edu.hk/~czhang22/pupsniffer-eval/Data/ccl2014_data.sql.

[4] The evaluation result on the Pupsniffer Evaluation Website, http://mega.ctl.cityu.edu.hk/~czhang22/pupsniffer-eval/result.jsp?recordtype=6.

(2) Error Analysis

According to Table 7, there are 409 false bilingual web pages pairs. We analyze these false pairs and classify them into four categories shown in Table 8.

Table 8. Types of Errors

Type of Error	Examples
Monolingual	www.family.org.hk/lang/en-us/carnival.html www.family.org.hk/lang/tc/carnival.html
Fake Bitext	www.tytaps.edu.hk/worksheet2/3A/English/8.pdf www.tytaps.edu.hk/worksheet2/3A/chinese/8.pdf
Error of the Content	www.chamber.org.hk/en/events/doc/M130519FF.pdf www.chamber.org.hk/tc/events/doc/M130519FF.pdf
Invalid Websites	www.polyu.edu.hk/fh/en-us/useful_links www.polyu.edu.hk/fh/tc/useful_links

I: Monolingual: The pairs URLs are in the same language.

II: Fake Bitext: The pairs URLs are false bitext according to their content.

III: Error of the content: One of the pages or both of the pages have been mentioned to have moved to other pages or not to exist in the certain websites any more.

IV: Invalid Websites: After the loading of the websites, it shows '404' or other hint that the web page is invalid.

Table 9 shows the distribution of false pairs. Nearly half of the incorrect pairs are due to the monolingual reason and another problem is the error of the content. If we can identify the language of the URL pairs, theoretically, we can filter out the monolingual URL pairs.

Table 9. False pairs distribution

Error Pairs	Error Type			
	Pairs of I	Pairs of II	Pairs of III	Pairs of IV
Number	204	40	126	39
Patio	49.88%	9.78%	30.81%	9.53%

(3) URL analysis

According to the 4460 pairs of URLs in the evaluation, we made an analysis about the site types of the URLs and the pattern distribution. Table 10 shows the distribution of different site types and their precisions.

Table 10. Distribution of different site types

Result	gov.hk	edu.hk	org.hk
TRUE	2 555	779	696
FALSE	133	180	96
Total	2 688	959	792
Precision	95.05%	81.23%	87.88%

From the Table 10, we can find that the precision of the governmental websites ranks the first in the three types. It reaches the precision of 95.05% which is absolutely above the total precision of our method. The other two types' precisions are all under 90%. It indicates that during our experiment, the governmental type of the web-

sites will be much more easily found compared with the educational websites and the organizational websites.

5 Conclusion and Future Works

In this paper we have presented a way to mine bilingual webpages with the help of search engine and the patterns replacing. When choosing the high credibility bilingual URL pairing patterns to do the replacement, we can find the corresponding Chinese URL in a fast way. The experiment ultimately collected a total of 153, 683 the English websites of the parallel sites, where there are 88 915 new URLs are determined to exist on the Internet. And the accuracy of actual parallel pages is 90.8%. Though the accuracy is not in accord with our expectations but there is still a big room for improvement.

In the future work, we plan to extract bilingual websites of other website types and search for the webpages of other districts like Taiwan, etc. We will also find the patterns with high credibility of different language pairs to see if the method still works on detecting the parallel websites of other languages.

Acknowledgments

This work is supported by National Natural Science Foundation of China through the grant (No.70903032), Major Projects of National Social Science Fund (13&ZD174), and National Social Science Fund Project (No.14BTQ033).

References

1. Oard, D.W.: Cross-language text retrieval research in the USA. In: Proceedings of the Third DELOS Workshop: Cross-Language Information Retrieval, pp. 7-16 (1997)
2. Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. Computational linguistics. 16(2) , pp. 79-85 (1990)
3. Resnik, P.: Parallel strands: A preliminary investigation into mining the web for bilingual text. In: D. Farwell, L. Gerber, and E. Hovy. (eds.) Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas (AMTA-98), pp. 72-82 (1998)
4. Resnik, P.: Mining the web for bilingual text. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 527-534 (1999)
5. Resnik, P., Smith, N.A.: The web as a parallel corpus. Computational Linguistics. 29(3), pp. 349-380 (2003)
6. Vicente, I. S., Manterola, I.: PaCo2: A Fully Automated tool for gathering Parallel Corpora from the Web. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 1-6 (2012)

7. Zhang, C. Z., Yao, X. C., Kit C. Y.: Finding More Bilingual Webpages with High Credibility via Link Analysis. In: Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pp. 138-143 (2013)
8. Ma, X., Liberman., M. Y.: Bits: a method for bilingual text search over the Web. In: Proceedings of MT Summit VII, pp. 13-17. Singapore (1999)
9. Zhang,Y., Wu, K., Gao J. F., Vines, P.: Automatic acquisition of Chinese-English parallel corpus from the web. In: Proceedings of 28th European Conference on Information Retrieval, pp. 420-431. London (2006)
10. Shi, L., Niu, C., Zhou, M., Gao, J. F.: A DOM tree alignment model for mining parallel data from the web. In: Proceedings of COLING/ACL-2006, pp. 489-496. Sydney (2006)
11. Miquel, E-G., Mikel, L. F.: Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with Bitextor. The Prague Bulletin of Mathematical Linguistics. No.93, pp. 77-86 (2010)
12. Ling, W., Xiang, G., Dyer, C., Black, A., Trancoso, I.: Microblogs as Parallel Corpora. In: The 51st Annual Meeting of the Association for Computational Linguistics (ACL) , pp. 176-186 (2013)
13. Ye, S.N., Lv, Y.J., Huang, Y., Liu, Q.: Automatic parallel sentences extracting from web. Journal of Chinese Information Processing. 22(5), pp. 67-73 (2008), (In Chinese)
14. Kit, C.Y., Ng, J.Y.H.: An Intelligent Web Agent to Mine Bilingual Parallel Pages via Automatic Discovery of URL Pairing Patterns. In: Proceedings of Web Intelligence and Intelligent Agent Technology Workshops, pp. 526-529 (2007)
15. Qi, L., Yang, L., Sun. M. S.: A Parallel Pages Mining Approach: Combing URL Patterns and HTML Structures. Journal of Chinese Information Processing. 27(3), pp. 91-99 (2013), (In Chinese)