Using HTML Tags to Improve Parallel Resources Extraction

Yan-hui Feng, Yu Hong, Wei Tang, Jian-min Yao, Qiao-ming Zhu
School of Computer Science & Technology, Soochow University, Suzhou, China
E-mail: hongy @ suda.edu.cn

***Abstract*—This paper proposes a new approach to extract parallel resources (including bilingual sentences and bilingual terms) from bilingual web pages, which have a primary language and a secondary language (the second language is often the translation to primary language). Our method is composed of four tasks: 1) parsing the web page into a DOM tree and segmenting inner texts of each node into series of monolingual snippets; 2) selecting adjacent snippet pairs in different languages and with higher translation scores as seeds for the next task; 3) constructing comprehensive wrappers from selected seeds, which save both HTML and surface formatting styles; 4) mining candidate instances and selecting good instances by their similarities with seeds. In this paper, we first propose to segment text by HTML tags, and select potential parallel resources by ranking all extracted candidates. According to the experimental results, our method can be applied to bilingual pages written in any other pair of languages. Experimental results also show that our approaches are effective in improving the parallel resources extraction.** *(Abstract)*

Keywords- Bilingual Resource, HTML Tags, Web Data Mining (key words)

I. INTRODUCTION

Parallel resources are critical for many NLP applications, such as machine translation [1] and cross language information retrieval [2]. Because it's hard to create large scale parallel dataset with human effort, many studies tried to extract parallel data automatically, such as from parallel monolingual web page pairs of some bilingual web sites [3] [4] [5] [6]. Candidate parallel web pages are acquired by making use of URL strings or HTML tags, then the translation equivalence of the candidate pairs are verified via content based features.

However, parallel resources exist not only in parallel monolingual web pages, but also in bilingual web pages. Statistically at least tens of millions of bilingual pages exist in Chinese web sites. People create such web pages for various reasons. The page[1] (see Figure 1) lists the most common oral sentences in English and their Chinese translation to facilitate English learning. Since such bilingual pages are very common in the Web, so it's an important task to mine parallel data from them. And in [7], an effective method has been proposed to acquire such bilingual web pages automatically by search engines.

Usually, the parallel data appears collectively and follow similar formatting styles in bilingual pages. Due to this phenomenon, in this paper, we propose a seed-expansion method to mine parallel data within a page. **Compared to the previous work, our method has the following advantages:**

- **HTML tags are used to improve parallel resources extraction.** The pages' physical structure such as HTML is far different from visual layout structure, and pages' editors arrange pages' content mainly from semantic. Superfluous to split contents with complete meaning into different tags. Former researchers all proposed to split text into different snippets by languages. In this paper, HTML tags are also seen as important clues to segment texts.
- **We propose to extract good candidate instances by ranking.** In [8], they focus on selecting good wrappers, and all candidate pairs extracted by them are regarded as parallel data. But they didn't pay attention to filter out noisy candidates extracted from good wrappers. Compared to [8], we propose to rank all candidates, extracted from all wrappers, by their relevance with seeds, but don't concern on the quality of wrappers.
- **Our method can be applied to bilingual web pages written in any pair of languages indiscriminately**, such as Japanese–English, Korean-English and so on, for that our approach is completely character-based and doesn't limit any language and domain.

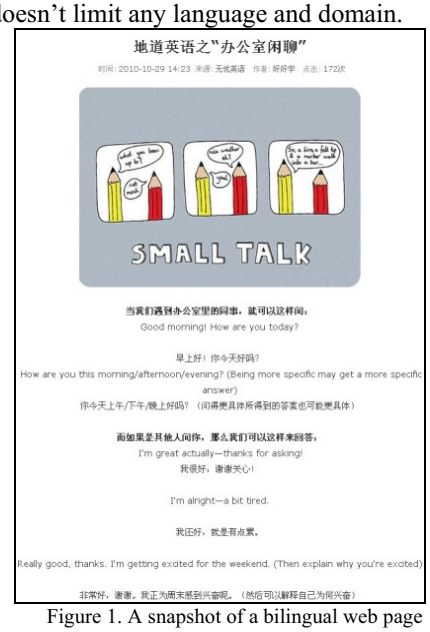

Figure 1. A snapshot of a bilingual web page

II. REALTED WORK

A. Parallel Data Mining from the Web

Most existing studies, like [2][4][6], mine parallel web documents within bilingual web sites first, then extract bilingual sentences from mined parallel documents. However, since the number of bilingual web sites is quite small, these methods can't yield a great many of parallel data. Besides, a method is proposed to discover bilingual sentences in comparable corpora in [9].

Based on that authors of pages often arrange terms with their translations inside a pair of parentheses, [10] and [11]

[1] http://english.51ielts.com/a/esp/office/2010/1029/30266.html

978-0-7695-4554-7/11 $26.00 © 2011 Crown Copyright
DOI 10.1109/IALP.2011.23

propose two different methods. However, since not all parenthesis patterns can be collected, these methods may miss a lot of translations. [12] [13] and [14] try to mine term translations from text snippet returned by search engines. They extract target translations by submitting the source term to search engines. It's difficult to mine low-frequency translations, for desired pages containing source and target translations can't be returned.

There are two publications available on automatically acquiring parallel data from bilingual web pages. An adaptive pattern-based method is used to mine interesting parallel data in [8]. They observe that many web pages contain parallel data collections, which follow a mostly consistent but possibly somewhat variable surface pattern. They select high-quality term pairs to learn candidate patterns first and then train a classifier to detect good patterns. All instances extracted from such good patterns are regarded as parallel data. However, even though a "good" pattern, it also can bring noise. In [15], they attempt the task of searching for high-quality term pairs from the Web. They formulate good search query using "Learning to Rank" and filter noisy document pairs with IBM Model 1 alignment.

III. OVERVIEW

As in Figure 2, our system is composed of four major components: the Page Analyzer, the Seed Extractor, the Wrapper Constructor and the Candidate Ranker.

For a given page, the Page Analyzer is proposed to get its DOM tree. Inner text of each node in DOM tree is segmented by HTML tags and languages. Each text snippet is term or sentence with complete meaning as well as just in one language.

Then high-quality translation pairs are identified by the Seed Extractor identifies as seeds for the following task. The Wrapper Constructor is used to learn wrappers automatically based on identified seeds. Wrappers in this paper save both seeds' surface and HTML formatting styles. With learnt wrappers, some potential parallel data, discarded by the seed mining module, can be extracted by simply matching.

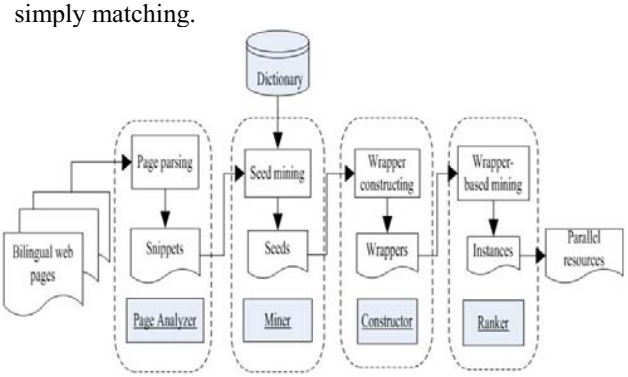

Figure 2. The workflow of the method

With the aim of extracting high-quality bilingual data, Candidate Ranker constructs a graph that models all the relations among seeds, wrappers, and candidate instances. Then it ranks all extracted candidates. Instances are ranked higher if they are related to seeds closely.

IV. PARALLEL RESOURCE EXTRACTION

A. The Page Analyzer

1) HTML Page Parsing

The Document Object Model (DOM) is an application programming interface used for parsing HTML documents. With DOM, an HTML document is parsed into a tree structure, when each node belongs to some predefined types (e.g. DIV, TABLE, TEXT, etc.). We removed nodes with types of "B", "I" and so on, because they are mainly used for controlling visual effect. After removal, their child nodes will be directly connected to their parents.

2) Text Segmentation

After a page is parsed, in our approach, the inner text of main data region should be segmented into a list of text snippets, such as $...C_iE_jC_{i+1}C_{i+2}C_{i+3}E_{j+1}E_{j+2}C_{i+4}E_{j+3}...$ (C and E stand for parallel snippets). Text snippets are expected to be the smallest unit with complete meaning. If text is only segmented by language, some semantically irrelevant text (but in the same language) would be merged together. Table Ⅰ shows four snippets of example page (in Figure 1) after segmenting by languages. The second snippet in Table Ⅰ contains two sentences, which are "*你今天上午/下午/晚上好吗？（问得更具体所得到的答案也可能更具体）*" and "*而如果是其他人问你，那么我们可以这样来回答*". The first sentence is the translation of the first snippet in Table Ⅰ and the second one is noisy. To acquire accurate bilingual snippet pairs, we need to segment snippets further by new clues.

Web pages are used to publish information on the Web, and web pages' editors always arrange a sentence under same tag. It's uncommon and unwise to segment sentences and terms into several parts and enwrap them respectively by different tags. According to that, HTML tags are used to segment text more accurately.

TABLE Ⅰ. EXAMPLE SNIPPETS SEGMENTED BY LANGUAGES

N	Snippets
1	How are you this morning/afternoon/evening? (Being more specific may get a more specific answer)
2	*你今天上午/下午/晚上好吗？（问得更具体所得到的答案也可能更具体）而如果是其他人问你，那么我们可以这样来回答：*
3	I'm great actually—thanks for asking!
4	*我很好, 谢谢关心！*

Part of the HTML source file is shown in Figure 3. Because of space limitation, we omit attribute values and only keep tags' names. According to it, the second snippet in Table Ⅰ could be segmented into two sentences by HTML tags "*<p> *". To decide the boundary of a snippet, two simple rules are defined as follows.

1) Open punctuation (like '(', '[') are added into next snippet; close punctuation (like ')', ']') are added into previous snippet; other punctuations (like ';') are added into previous snippet.

2) Abbreviations (like 'IALP') are merged into previous and next snippets in different languages.

B. The Seed Extractor

To select high-quality parallel snippet pairs as seeds for next work, a length-based model and an alignment model are adopted in this part. Adjacent snippets, which are in different languages and match the two models, could be selected as seeds for wrapper's construction.

1) Length-based Model

In [16], the length model is based on the assumption that each word in one language, l_1, gives rise to a random number of words in the other language, l_2. They assume these random variables are independent and distributed identically with a normal distribution. The model is then specified by the mean, c, and variance, s^2, of this distribution, c is the expected number of characters in l_2 per character in l_1 and s^2 is the variance of the number of characters in l_2 per character in l_1. They define:

$$\sigma = (l_2 - l_1 c) \Big/ \sqrt{l_1 s^2} \qquad (1)$$

We calculate c and s^2 with 6,500 pairs of Chinese-English sentences, and get the value 1.03 and 0.426 respectively and get the σ threshold between 4.722~15.27.

2) Alignment Model

Word overlap judges the similarity of bilingual snippets. In this paper, we acquire the word-overlap score between any two adjacent language segments. The similarity *Score(c_res,e_res)* of bilingual snippets is based on word-overlap as following:

$$Score(c\_res, e\_res) = \frac{\sum_{i=1}^{p} Max_{1 \le j \le q}(Sim(c_i, e_j))}{\phi} \qquad (2)$$

where p and q stand for the lengths of snippets in source and target languages. The denominator is normalization factor, and in our experiment we select $p+q$ as its value. In addition, c_i stands for the i^{th} word of Chinese term and e_j stands for the j^{th} word of English term. $Sim(c_i,e_j)$ in [17] stands for the similarity of c_i and e_j.

C. The Wrapper Constructor

Many parallel contents, which are discarded by the seed extractor, may be mined by learnt wrappers. In this section, we will describe an approach to construct wrappers automatically. Information of seeds' surface and HTML formatting style is saved in our wrappers.

1) The Main Algorithm

First, the surface formatting style is related to seed's content; HTML formatting style is related to HTML tags enwrapped seed. Both of them are character-based rules. Each seed is a pair of bilingual snippets.

For each seed, the method in [8] is adopted to get its surface formatting styles. The templates are learnt based on all processed strings. For example, "*I'm alright—a bit tired.*" and "*我还好，就是有点累。*" are selected as a seed, then we can learn a template, which are "*[E][C]*".

For acquiring HTML formatting style of each seed, let s_i be the i^{th} seed, l_i, and r_i, be series of HTML tags preceding and following s_i. Let m_i be the series of tags between parallel contents. And m_i is null when no tag exists between them. For the seed "*I'm alright—a bit tired.*" and "*我还好，就是有点累。*", according to the HTML source codes in Figure 3, "*
 " and "
 " are the HTML tag sequences preceding and following it. And the middle tags between the two snippets are "
 *".

```
… <div><h2>地道英语之“办公室闲聊” </h2></div>
    <div><small>时间:</small>2010-10-29 14:23…</div>
<div class="content"><table width='100%'><tr><td><center><img/><br />
<span class="story_cn txt-14">
<p><strong>当我们遇到办公室里的同事，就可以这样问： </strong></p>
<div class="langs_en">Good morning! How are you today?<br /></div>
<div class="langs_cn">早上好！你今天好吗？ </div>
<p>How are you this morning/afternoon/evening? (Being more specific may get a more specific answer)</p>
<div class="langs_cn">你今天上午/下午/晚上好吗？（问得更具体所得到的答案也可能更具体） <br /></div>
<p><strong>而如果是其他人问你，那么我们可以这样来回答： </strong></p>
<p>I'm great actually-thanks for asking!</p>
<div class="langs_cn">我很好，谢谢关心！ <br /></div>
<div class="langs_en">I'm alright-a bit tired.<br /></div>
<div class="langs_cn">我还好，就是有点累。<br /></div>
<div class="langs_en">Really good, thanks. I'm getting excited for the weekend. (Then explain why you're excited)<br /></div>
<div class="langs_cn">非常好，谢谢。我正为周末感到兴奋呢。（然后可以解释自己为何兴奋） <br /></div>
<div class="langs_en">Can't complain!<br /></div>
<div class="langs_cn">好得没话说！ <br /></div>
<p><strong>回答完后还可以问上一句： </strong></p>
<div class="langs_en">How about yourself?<br /></div>
<div class="langs_cn">你如何呢? </div>
</span></center></td></tr></table></div>
… …
```

Figure 3. HTML source file of the example page

2) An Example

We show an example of the wrapper construction. Three bilingual snippet pairs (i.e. S_1, S_2, and S_3 in Table Ⅱ) are selected as seeds from the example page (in Figure 1). Table Ⅱ lists selected seeds, constructed wrappers and extracted candidate instances.

TABLE Ⅱ. WRAPPERS CONSTRUCTING FROM EXAMPLE PAGE

S_1: *Really good,… excited) 非常好，…兴奋）* ***Wrapper constructed from S_1:*** * [E] [C] * ***Extractions:*** *I'm alright-a bit tired. 我还好，就是有点累。* *Can't complain! 好得没话说！*
S_2: *How about yourself? 你如何呢？* ***Wrapper constructed from S_2:*** * [E] [C] * ***Extractions:*** *Good morning! How are you today?* *早上好！你今天好吗？*
S_3: *I'm great … asking! 我很好，谢谢关心！* ***Wrapper constructed from S_3:*** * [E] [C] * ***Extractions:*** *How are you this morning/ … specific answer)* *你今天上午/ … 具体）*

To show the advantage of HTML formatting styles in avoiding noisy data, another example page is given (in Figure 4). "*新闻组 News*" and "*News Groups*" are selected as seeds, the learnt wrapper is *"<FONT color=#333399>[C][E]
 </FONT>*". Within the whole page only six non-noisy instances can be extracted, such as "*电子邮件 E-mail*" and "*Electronic Mail*". However, if wrappers don't contain HTML formatting

styles (as the wrappers in [8]), the wrapper should be "*[C][E]*". Based on this wrapper, all the adjacent snippets in different languages would be extracted as translations. Thus, too much noise is collected, like "*webmaster@chinadzz.com*" and "*服务类型是电子邮件， 收件人的地址是*" are seen as translations. Using HTML tags, such noise can be avoided greatly.

Figure 4. A Snapshot of a bilingual page

D. Candidates Ranker

In order to get the relevance between extracted candidates and seeds, we first need to understand how they are related globally. For one given page,

- Different wrappers can be learnt from one seed;
- The same wrapper can be learnt from different seeds;
- Same candidate can be extracted by several wrappers.

To model these complex relations, a directed graph, contains all the objects of page, seeds, wrappers, and extracted bilingual snippets and models all the relations between them, can be established. And the relevance between nodes is used to rank extracted candidates.

The first column of Table Ⅲ shows all possible node types, the middle column reports each of their possible relations. Target node types are shown in the last column. We assume that there are no edges from a node to itself.

TABLE Ⅲ. NODE AND RELATION TYPE

Source Type	Edge Relation	Target Type
page	find	seeds
seeds	derive find^{-1}	wrappers page
wrappers	extract derive^{-1}	candidates seeds
candidates	extract^{-1}	wrappers

How closely related are two nodes in a graph? The Candidate Ranker performs Random Walk with Restart [18], which provides a good relevance score between two nodes in a weighted graph, and then ranks nodes by their final score. It is defined as follows: consider a random particle that starts from node *x*. The particle iteratively transmits to its neighborhood with the probability that is proportional to their edge weights. Also at each step, it has some probability to return to node *x*. The relevance score of node *y* with node *x* is defined as the steady-state probability that the particle will finally stay at node *y*.

In [18], nodes are weighted higher if they are connected to many seed nodes by many short, low fan-out paths. The ranker stops until all node weights converge. In this paper, the initial "restart" set is the set of seeds.

V. EXPERIMENTS AND RESULTS

In this section, we will report some experimental results on two gold standard datasets.

A. Parallel Resources Extraction on Chinese-English Bilinglual Web Pages

1000 Chinese-English bilingual web pages are chosen from 12 popular Chinese web sites randomly, and all bilingual data are annotated manually. In this paper, the extracted data are considered correct only if they are exactly same as the data labeled by human.

To measure the improvement of using HTML tags to extract parallel data, we produce three alternative systems.

1) Baseline System: It is developed according to the method in [8], which doesn't take HTML tags into consideration. According to the results shown in Table Ⅳ, it gets 75.753% F-score. Compared to [8], where the F-score is 79.9%, it's obvious that our baseline system is an effective reproduce of the method in [8].

2) Ranker-based System: To filter noises in candidate translations, in this paper, candidates are ranked by their relevancies with seeds and wrappers. Thus we develop the ranker-based system. Compared to baseline system, it doesn't train a classifier to detect good wrappers. It tries to ranks candidate instances extracted by all wrappers and then select top ones.

3) HTML Tags & Ranker-based System: It is the full implementation of the proposed approach in this paper.

TABLE Ⅳ. PERFORMANCE OF DIFFERENT SYSTEMS BASED ON CHINESE-ENGLISH BILINGUAL PAGES

	P(%)	R(%)	F-score(%)
Baseline	70.2	82.26	75.753
Ranker-based	72.6	84.9	78.26
HTML Tags & Ranker	80.23	88.31	84.07

The experimental results are shown in Table Ⅳ, which verifies our method improve precision greatly. The main reason is that most of the noisy instances extracted by baseline system are avoided by use of HTML tags (the example mentioned in the second part of Section C properly verifies that). Besides that, the improvement also comes from the following aspects:

- Contents are segmented wrongly only by languages. Most of them contain some noisy text and therefore they are not same as the data labeled by human. Using HTML tags, they can be segmented more accurately;
- The translations (extracted by so-called bad wrappers in [8]) aren't ignored in ranking process. In [8], all candidates extracted from bad wrappers are ignored;
- The noisy translations (extracted by good wrappers in [8]) are ranked lower in our method because of loose

relevance with seeds. In [8], all candidate instances extracted good wrappers are collected as good ones.

Table Ⅳ also shows that the recall improves limitedly. The improvement mainly comes from:

- The translations are segmented correctly by use of HTML tags;
- The few high-quality translations (extracted from bad wrappers in [8]) are not discarded.

B. Parallel Resources Extraction on Japanese-English Bilinglual Web Pages

We also experiment on 1000 Japanese-English bilingual pages. Totally 3518 parallel translations are annotated. Table Ⅴ gives the result. Due to that Japanese always use some emoticons when writing, the precision value, recall value and F-measure value is lower than those of Chinese-English resource extraction. However, their extraction performances are also comparable.

TABLE Ⅴ. PERFORMANCE OF JAPANESE-ENGLISH RESOURCES EXTRACTION

	P(%)	R(%)	F-score(%)
Baseline	69.54	78.4	73.7
Ranker-based	71.27	80.18	75.46
HTML Tags & Ranker	78.86	82.4	80.59

C. Discussion

In the module of selecting seeds, some sentence pairs, which are not translations to each other (one sentence just is explanative texts of the other sentence), are always selected as seeds wrongly because of higher translation score. Such sentences always appear in language learning web site, with the aim of telling the learners relative grammar. To improve performance, we also need better methods to avoid the bad impact of such noisy seeds.

VI. CONCLUSION

Bilingual web pages have shown great potential as a source of up-to-date parallel data. This paper presents a novel method to extract parallel resources automatically from bilingual web pages. Based on the observation that parallel resources usually have similar surface and HTML formatting style within a page, a method through seed expansion is proposed. In this paper, we first propose to use HTML tags to improve accuracy of text segmentation, and we first propose to rank candidate instances to select high-quality ones. Our method is page-, domain- and language- independent. According to experimental results on two manually made test data sets, our method is quite promising.

As a valuable resource for many NLP applications, such as machine translation and cross language information retrieval, our method brings an efficient and effective solution to bilingual language engineering. In the future, we want to evaluate the usefulness of our mined data for machine translation or other applications.

ACKNOWLEDGMENTS

We acknowledge the support of the National Natural Science Foundation of China under Grant No. 60970057, 61003152, and Municipal Foundation SYG201030.

REFERENCES

[1] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19:2, pp: 263-311. 1993.

[2] J-Y Nie, M. Simard, P. Isabelle, and R. Durand. Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of parallel Text from the Web. In *SIGIR*. pp: 74-81. 1999.

[3] Jiang Chen and Jian-Yun Nie. Web Parallel text mining for Chinese-English cross-language information retrieval. *Proceedings of RIAO2000 Content-Based Multimedia Information Access, CID, Paris*. 2000.

[4] Philip Resnik and Noah A. Smith. The web as a Parallel Corpus. *Computational Linguistics*. 2003.

[5] Ying Zhang, Ke Wu, Jianfeng Gao, Phil Vines. Automatic Acquisition of Chinese-English Parallel Corpus from the web. In *Proceedings of 28th European Conference on Information Retrieval*. 2006.

[6] Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. A DOM Tree Alignment Model for Mining Parallel Data from the Web. In *Joint Proceedings of the Association for Computa-tional Linguistics and the International Conference on Computational Linguistics*, Sydney, Australia. 2006.

[7] Yanhui Feng, Yu Hong, Zhenxiang Yan, Jianmin Yao, Qiaoming Zhu. 2010. A Novel Method for Bilingual Web Page Acquisition from Search Engine Web Records. In *COLING 2010*, Poster Volume, pp: 294–302.

[8] Long Jiang, Shiquan Yang, Ming Zhou, Xiaohua Liu and Qingsheng Zhou. Mining Bilingual Data from the Web with Adaptively Learnt Patterns. *The 47th Annual Meeting of the Association for Computational Linguistics*. pp: 870-878. 2009.

[9] D. S. Munteanu, D. Marcu. Improving Machine Translation Performance by Exploiting Non- Parallel Corpora. *Computational Linguistics*. 31(4). pp: 477-504. 2005.

[10] G.H. Cao, J.F. Gao and J.Y. Nie. A system to mine large-scale bilingual dictionaries from monolingual web pages. *MT summit*. pp: 57-64. 2007.

[11] D. Lin, S. Zhao, B. Durme and M. Pasca. Mining Parenthetical Translations from the Web by Word Alignment. In *ACL*. pp: 994-1002. 2008.

[12] Ying Zhang, Phil Vines. Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In the *Proceedings of SIGIR 2004, pp:162-169.*

[13] Pu-Jen Cheng , Jei-Wen Teng , Ruei-Cheng Chen , Jenq-Haur Wang , Wen-Hsiang Lu , Lee-Feng Chien. Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval. In the *Proceedings of SIGIR*, 2004.

[14] Fei Huang, Ying Zhang, Stephan Vogel, Mining key phrase translations from web corpora, In the *Proceedings of HLT-EMNLP*, pp: 483-490. 2005.

[15] Gumwon Hong, Chi-Ho Li, Ming Zhou and Hae-Chang Rim. An Empirical Study on Web Mining of Parallel Data. In *COLING 2010*, pp: 474-482.

[16] William A. Gale, Kenneth W. Church, A program for aligning sentences in bilingual corpora, *Computational Linguistics*, v.19 n.1, March 1993.

[17] Feifan Liu, Jun Zhao, Bo Xu. Building Large-Scale Domain Independent Chinese English Bilingual Corpus and the Researches on Sentence Alignment. *Joint Symposium on Computational Linguistics*. 2003.

[18] H. Tong, C. Faloutsos, and J.-Y. Pan. Random walk with Restart: Fast solutions and applications. Knowledge and Information Systems: An International Journal (KAIS), 2007.