An Iterative Link-based Method for Parallel Web Page Mining

Le Liu[1], Yu Hong[1], Jun Lu[2], Jun Lang[2], Heng Ji[3], Jianmin Yao[1]
[1]School of Computer Science & Technology, Soochow University, Suzhou, 215006, China
[2]Institute for Infocomm Research, Singapore, 138632
[3]Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
giden@sina.cn,{tianxianer,lujun59,billlangjun}@gmail.com
jih@rpi.edu,jyao@suda.edu.cn

Abstracts

Identifying parallel web pages from bilingual web sites is a crucial step of bilingual resource construction for cross-lingual information processing. In this paper, we propose a link-based approach to distinguish parallel web pages from bilingual web sites. Compared with the existing methods, which only employ the internal translation similarity (such as content-based similarity and page structural similarity), we hypothesize that the external translation similarity is an effective feature to identify parallel web pages. Within a bilingual web site, web pages are interconnected by hyperlinks. The basic idea of our method is that the translation similarity of two pages can be inferred from their neighbor pages, which can be adopted as an important source of external similarity. Thus, the translation similarity of page pairs will influence each other. An iterative algorithm is developed to estimate the external translation similarity and the final translation similarity. Both internal and external similarity measures are combined in the iterative algorithm. Experiments on six bilingual websites demonstrate that our method is effective and obtains significant improvement (6.2% F-Score) over the baseline which only utilizes internal translation similarity.

1 Introduction

Parallel corpora have played an important role in multilingual Natural Language Processing, especially in Machine Translation (MT) and Cross-lingual Information Retrieval(CLIR). However, it's time-consuming to build parallel corpora manually. Some existing parallel corpora are subject to subscription or license fee and thus not freely available, while others are domain-specific. Therefore, a lot of previous research has focused on automatically mining parallel corpora from the web.

In the past decade, there have been extensive studies on parallel resource extraction from the web (e.g., Chen and Nie, 2000; Resnik 2003; Jiang et al., 2009) and many effective Web mining systems have been developed such as STRAND, PTMiner, BITS and WPDE. For most of these mining systems, there is a typical parallel resource mining strategy which involves three steps: (1) locate the bilingual websites (2) identify parallel web pages from these bilingual websites and (3) extract bilingual resources from the parallel web pages.

In this paper, we focus on the step (2) which is regarded as the core of the mining system (Chunyu, 2007). Estimating the translation similarity of two pages is the most basic and key problem in this step. Previous approaches have tried to tackle this problem by using the information within the pages. For example, in the STRAND and PTMiner system, a structural filtering process that relies on the analysis of the underlying HTML structure of pages is used to determine a set of pair-specific structural values, and then the values are used to decide whether the pages are translations of one another. The BITS system filters out bad pairs by using a large bilingual dictionary to compute a content-based similarity score and comparing the score with a threshold. The WPDE system combines URL similarity, structure similarity with content-based similarity to discover and verify candidate parallel page pairs. Some other features or rules such as page size ratio, predefined hypertexts which link to different language versions of a web page are also used in most of these systems. Here, all of the mining systems are simply using the information within the page in the process of find-

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1216–1224,
October 25-29, 2014, Doha, Qatar.

ing parallel web pages. In this paper, we attempt to explore other information to identify parallel web pages.

On the Internet, most web pages are linked by hyperlinks. We argue that the translation similarity of two pages depends on not only their internal information but also their neighbors. The neighbors of a web page are a set of pages, which link to the page. We find that the similarity of neighbors can provide more reliable evidence in estimating the translation similarity of two pages.

The main issues are discussed in this paper as follows:

- *Can the neighbors of candidate page pairs really contribute to estimating the translation similarity?*
- *How to estimate the translation similarity of candidate page pairs by using their neighbors?*

Our method has the following advantages:

High performance

The external and internal information is combined to verify parallel page pairs in our method, while in previous mining systems, only internal information was used. Experimental results show that compared with existing parallel page pair identification technologies, our method obtains both higher precision and recall (6.2% and 6.3% improvement than the baseline, respectively). In addition, the external information used in our method is a more effective feature than internal features alone such as structural similarity and content-based similarity.

Language independent

In principle, our method is language independent and can be easily ported to new language pairs, except for the language-specific bilingual lexicons. Our method takes full advantage of the link information that is language-independent. For the bilingual lexicons in our experiments, compared to previous methods, our method does not need a big bilingual lexicon, which is good news to less-resource language pairs.

Unsupervised and fewer parameters

In previous work, some parameters need to be optimized. Due to the diversity of web page styles, it is not trivial to obtain the best parameters. Some previous researches(Resnik, 2003; Zhang et al., 2006) attempt to optimize parameters by employing machine learning method. In contrast, in our method, only two parameters need to be estimated. One parameter remains stable for different style websites. Another parameter can be easily adjusted to achieve the best performance. Therefore, our method can be used in other websites with different styles, without much effort to optimize these parameters.

2 Related Work

A large amount of literature has been published on parallel resource mining from the web. According to the existing form of the parallel resource on the Internet, related work can be categorized as follows:

Mining from bilingual websites

Most existing web mining systems aimed at mining bilingual resource from the bilingual websites, such as PTMiner (Nie et al., 1999), STRAND (Resnik and Smith, 2003), BITS (Ma and Liberman, 1999), PTI (Chen et al., 2004). PTMiner uses search engines to pinpoint the candidate sites that are likely to contain parallel pages, and then uses the collected URLs as seeds to further crawl each web site for more URLs. Web page pairs are extracted based on manually defined URL pattern matching, and further filtered according to several criteria. STRAND uses a search engine to search for multilingual websites and generated candidate page pairs based on manually created substitution rules. Then, it filters some candidate pairs by analyzing the HTML pages. PTI crawls the web to fetch (potentially parallel) candidate multilingual web documents by using a web spider. To determine the parallelism between potential document pairs, a filename comparison module is used to check filename resemblance, and a content analysis module is used to measure the semantic similarity. BITS was the first to obtain bilingual websites by employing a language identification module, and then for each bilingual website, it extracts parallel pages based on their content.

Mining from bilingual web pages

Parallel/bilingual resources may exist not only in two parallel monolingual web pages, but also in single bilingual web pages. Jiang et al. (2009) used an adaptive pattern-based method to mine interesting bilingual data based on the observation that bilingual data usually appears collectively following similar patterns. They found that bilingual web pages are a promising source of up-to-date bilingual terms/sentences which cover many domains and application scenarios. In addition, Feng et al. (2010) proposed a new method

to automatically acquire bilingual web pages from the result pages of a search engine.

Mining from comparable corpus

Several attempts have been made to extract parallel resources from comparable corpora. Zhao et al. (2002) proposed a robust, adaptive approach for mining parallel sentences from a bilingual comparable news collection. In their method, sentence length models and lexicon-based models were combined under a maximum likelihood criterion. Smith et al. (2010) found that Wikipedia contains a lot of comparable documents, and adopted a ranking model to select parallel sentence pairs from comparable documents. Bharadwaj et al. (2011) used a SVM classifier with some new features to identify parallel sentences from Wikipedia.

3 Iterative Link-based Parallel Web Pages Mining

As mentioned, the basic idea of our method is that the similarity of two pages can be inferred from their neighbors. This idea is illustrated in Figure 1.

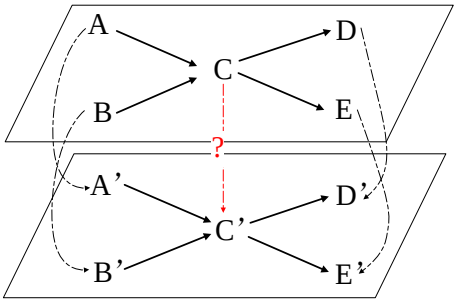

Figure 1 Illustration of the link-based method

In Figure 1, *A*, *B*, *C*, *D* and *E* are some pages in the same language; while *A'*, *B'*, *C'*, *D'* and *E'* are some pages in another language. The solid black arrows indicate the links between these pages. For example, page *A* points to *C*, page *B'* points to *C'* and so on. Then the page set {*A*, *B*, *D*, *E*} is called the neighbors of page *C*. Similarly, the page set {*A'*, *B'*, *D'*, *E'*} contains the neighbors of page *C'*. If the page pairs : <*A*, *A'*>, <*B*, *B'*>, <*D*, *D'*> and <*E*, *E'*> have high translation similarities, then it can be inferred that page *C* and *C'* have a high probability to be a pair of parallel pages. Every page has its own neighbors. For each web page, our method views link-in and link-out hyperlinks as the same. Thus, the linked pages will influence each other in estimating the translation similarity. For example, the similarities of two pairs <*A*, *A'*> and <*C*, *C'*> will influence each other. It is an iterative process. We will elaborate the process in the following sections.

Since our goal is to find parallel pages in a specific website, the key task is to evaluate the translation similarity of two pages (which are in different languages) as accurately as possible. The final similarity of two pages should depend both on their internal similarity and external similarity. The internal similarity means the similarity estimated by using the information in the page itself, such as the structure similarity and the content-based similarity of the two pages. On the other hand, the external similarity of two pages is the similarity depending on their neighbors. The final translation similarity is called the **Enhanced Translation Similarity** (*ETS*). The *ETS* of two pages can be calculated as follows:

$$ETS(e,c) = \alpha \cdot S_{ext}(e,c) + (1-\alpha) \cdot S_{in}(e,c), \alpha \in [0,1] \quad (1)$$

Where, $S_{in}(e,c)$ is the internal translation similarity of two pages: e and c; $S_{ext}(e,c)$ represents the external translation similarity of pages e and c. $ETS(e,c)$ indicates the final similarity of two pages, which combines the internal with external translation similarity.

In this paper, we conduct the experiments on English-Chinese parallel page pair mining. However, our method is language-independent. Thus, it can be applied to other language pairs by only replacing a bilingual lexicon. The symbol e and c always indicate an English page and a Chinese page respectively in this paper. In the following sections, we will describe how to calculate the $S_{in}(e,c)$ and $S_{ext}(e,c)$ step by step.

3.1 Preprocessing

The input of our method is a bilingual website. This paper aims to find English/Chinese parallel pages. So a 3-gram language model is used to identify (or classify) the language of a certain document. The performance of the language identification module achieves 99.5% accuracy through in-house testing. As a result, a set of English pages and a set of Chinese pages are obtained. In order to get the neighbors of a page, for each bilingual website, two networks are constructed based on the hyperlinks, one for English pages and another for Chinese pages.

3.2 The Internal Translation Similarity

Following Resnik and Smith (2003), three features are used to evaluate the internal translation similarity of two pages:

The size ratio of two pages

The length ratio of two documents is the simplest criterion for determining whether two documents are parallel or not. Parallel documents tend to be similar in length. And it is reasonable to assume that for text *E* in one language and text *F* in another language, length(*E*) ≈ *C*•length(*F*), where *C* is a constant that depends on the language pair. Here, the content length of a web page is regarded as its length.

The structure similarity of two pages

The HTML tags describe and control a web page's structure. Therefore, the structure similarity of two pages can be calculated by their HTML tags. Here, the HTML tags of each page are extracted (except the visual tags such as "B", "FONT".) as a linear sequence. Then the structure similarity of two pages is computed by comparing their linearized sequences. In this paper, the LCS algorithm (Dan, 1997) is adopted to find the longest common sequences of the two HTML tag sequences. The ratio of LCS length and the average length of two HTML tag sequences are used as the structure similarity of the two pages.

The content-based translation similarity of two pages

The basic idea is that if two documents are parallel, they will contain word pairs that are mutual translations (Ma, 1999). So the percentage of translation word pairs in the two pages can be considered as the content-based similarity. The translation words of two documents can be extracted by using a bilingual lexicon. Here, for each word in English document, we will try to find a corresponding word in Chinese document.

Finally, the internal translation similarity of two pages is calculated as follows:

$$S_{in}(e,c) = \beta \cdot S_{cb}(e,c) + (1-\beta) \cdot S_{struct}(e,c), \beta \in [0,1] \quad (2)$$

Where, $S_{cb}(e,c)$ and $S_{struct}(e,c)$ are the content-based and structural similarity of page e and c respectively. In addition, the size ratio of two pages is used to filter invalid page pairs.

3.3 The External and Enhanced Translation Similarity

As described above, the external translation similarity of two pages depends on their neighbors:

$$S_{ext}(e,c) = Sim\big(PG(e), PG(c)\big) \quad (3)$$

Where, *PG(x)*, a set of pages, is the neighbors of page *x*. Obviously, the similarity of two sets relies on the similarity of the elements in the two sets. Here, the elements are namely web pages. So, $S_{ext}(e,c)$ equals to $Sim\big(PG(e), PG(c)\big)$, and $Sim\big(PG(e), PG(c)\big)$ depends on $ETS\big(e_i, c_j\big)$ (e_i, c_j belongs to $PG(e)$, $PG(c)$, respectively) and $ETS(e,c)$. According to Equation (1), $ETS(e,c)$ depends on $S_{in}(e,c)$ and $S_{ext}(e,c)$. Therefore, it is a process of iteration. $ETS(e,c)$ will converge after a certain number of iterations. Thus, $ETS^i(e,c)$ is defined as the enhanced similarity of page *e* and *c* after the *i*-th iteration, and the same is for $S^i_{ext}(e,c)$ and $Sim^i\big(PG(e), PG(c)\big)$. $Sim^i\big(PG(e), PG(c)\big)$ is computed by the following algorithm:

Algorithm 1: Estimating the external translation similarity

Input: $PG(e), PG(c)$
Output: $S^i_{ext}(e,c)$

Procedure:

$sum \leftarrow 0$
$e\_set \leftarrow PG(e)$
$c\_set \leftarrow PG(c)$
While e_set and c_set are both not empty:
  $< x, y >$
  $\leftarrow arg\,max_{x \in e\_set,\ y \in c\_set}\ (ETS^{i-1}(x,y))$
  $sum \leftarrow sum +$ $\mathrm{ETS}^{i-1}(x,y)$
  Remove x from e_set
  Remove y from c_set
$S^i_{ext}(e,c) = Sim^i\big(p(e), p(c)\big)$
$= 2 \cdot sum \,/\, (|PG(e)| + |PG(c)|)$

Algorithm 2 Estimating the enhanced translation similarity

Input: P_e, P_c, (*the English and Chinese page set*)
Output: ETS(e, c), $e \in P_e, c \in P_c$
Initialization: Set *ETS*(*e*, *c*) random value or small value

Procedure:

LOOP:
  For each e in P_e :
    For each c in P_c:
      $ETS^i(e,c) = \alpha \cdot S^i_{ext}(e,c) + (1-\alpha) \cdot S_{in}(e,c)$
  Parameters normalization
UNTIL $ETS(e,c)$ is stable

Algorithm 1 tries to find the real parallel pairs from $PG(e)$ and $PG(c)$. The similarity of $PG(e)$ and $PG(c)$ is calculated based on the similarity

values of these pairs. Finally, $ETS(e,c)$ is calculated by the following algorithm 2.

In Algorithm 2, the input P_e and P_c are English and Chinese page sets in a certain bilingual website. We use algorithm 2 to estimate the enhanced translation similarity.

3.4 Find the Parallel Page Pairs

At last, the enhanced translation similarity of every pair is obtained, and the parallel page pairs can be extracted in terms of these similarities:

Algorithm 3 Finding parallel page pairs

Input: P_e, P_c
$ETS(x,y),\ \ x \in P_e, y \in P_c$
MAX_P (or MIN_SIM)

Output: Parallel Page Pairs List : PPL

Procedure:

LOOP:
$< x,y > = \ arg\,max_{x \in P_e, y \in P_c}\,(ETS(x,y))$
Add $< x,y >$ to PPL
Remove x from P_e
Remove y from P_c
UNTIL size of PPL > MAX_P (or $ETS(x,y)$ < MIN_SIM)

This algorithm is similar to Algorithm 1 in each bilingual website. The input MAX_P is an integer threshold which means that only top MAX_P page pairs will be extracted in a certain website. It needs to be noted that MAX_P is always less than $|P_e|$ and $|P_c|$. While the input MIN_SIM is another kind of threshold that is used for extracting page pairs with high translation similarity.

4 Experiments and Analysis

4.1 Experimental setup

Our experiments focus on six bilingual websites. Most of them are selected from HK government websites. All the web pages were retrieved by using a web site download tool: HTTrack[1]. We notice that a small amount of pages doesn't always contain valuable contents. So, we put a threshold (100 bytes in our experiment) on the web pages' content to filter meaningless pages. In order to evaluate our method, the bilingual page pairs of each website are annotated by a human annotator. Finally, we got 23109 pages and 11684 bilingual page pairs in total for testing. The basic information of these websites is listed in Table 1.

It's time-consuming to annotate whether two pages is parallel or not. Note that if a website contains *N* English pages and *M* Chinese pages, an annotator has to label *N*M* page pairs. To the best of our knowledge, there is no large scale and public parallel page pair dataset with human annotation. So we try to build a reliable and large-scale dataset.

In our experiments, *URL* similarity is used to reduce the workload for annotation. For a certain website, firstly, we obtain its *URL* pattern between English and Chinese pages manually. For example, in the website "www.gov.hk", the *URL* pairs like:

http://www.gov.hk/en/about/govdirectory/ (English)

http://www.gov.hk/sc/about/govdirectory/ (Chinese)

The *URL* pairs always point to a pair of parallel pages. So <"/en/","/sc/"> is considered as a *URL* pattern that was used to find parallel pages. For the other *URLs* that can't match the pattern, we have to label them by hand. The column "No pattern pairs" in Table 1 shows that the number of parallel page pairs which mismatch any patterns.

Table 1 Number of pages and bilingual page pairs of each websites

Site ID	En/Ch pages	Total pairs	No pattern pairs	URL
S1	1101/1098	1092	20	www.gov.hk
S2	501/497	487	7	www.customs.gov.hk
S3	995/775	768	12	www.sbc.edu.sg
S4	4085/3838	3648	4	www.swd.gov.hk
S5	660/637	637	0	www.landsd.gov.hk
S6	4733/4626	4615	8	www.td.gov.hk
total	12075/11471	11684	51	

Each website listed in Table 1 has a *URL* pattern for most parallel web pages. Some previous researches used the *URL* similarity or patterns to find parallel page pairs. However, due to the diversity of web page styles and website maintenance mechanisms, bilingual websites adopt varied naming schemes for parallel documents (Shi, et al, 2006). The effect of *URL* pattern-based mining always depends on the style of website. In order to build a large dataset, the *URL* pattern is not used in our method. Our method is able to handle bilingual websites without *URL* pattern rules.

In addition, an English-Chinese dictionary with 64K words pairs is used in our experiments. Algorithm 3 needs a threshold MAX_P or

[1] http://www.httrack.com/

MIN_SIM. It is very hard to tune the MIN_SIM because it varies a lot in different websites and language pairs. However, Table 1 shows that the number of parallel pages is smaller than that of English and Chinese pages. Here, for each website, the MAX_P is set to the number of Chinese pages (which is always smaller than that of English pages). In this way, the precision will never reach 100%, but it is more practical in a real application. As a result, in some experiments, we only report the F-score, and the precision and recall can be calculated as follows:

$$Precision = \frac{F_{score} \cdot (N_{Pairs} + MAX\_P)}{2 \cdot MAX\_P} \quad (4)$$

$$Recall = \frac{F_{score} \cdot (N_{Pairs} + MAX\_P)}{2 \cdot N_{Pairs}} \quad (5)$$

Where, N_{Pairs} for each website is listed in the "Total pairs" column of Table 1.

4.2 Results and Analysis

Performance of the Baseline

Let's start by presenting the performance of a baseline method as follows. The baseline only employs the internal translation similarity for parallel web pages mining. Algorithm 3 is also used to get the page pairs in baseline system. Here, the input $ETS(x, y)$ is replaced by $S_{in}(x, y)$. The parameter β in Equation 2 is a discount factor. For different β values, the performance of baseline system on six websites is shown in Figure 2. In the Figure 2, it shows that when β is set to 0.6, the baseline system achieves the best performance. The precision, recall and F-score are 85.84%, 87.55% and 86.69% respectively. So in the following experiments, we always set β to 0.6.

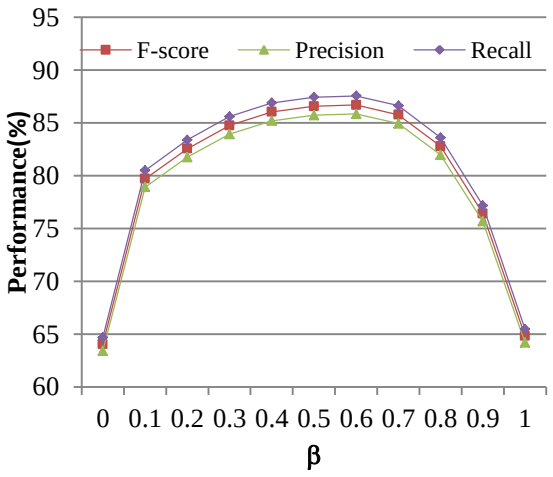

Figure 2 Performances of baseline system with different β value

Performance of Our Method

As described in Section 3, our method combines the internal with external translation similarity in estimating the final translation similarity (i.e., ETS) of two pages. So, the discount factor α in Equation (1) is important in our method. Besides, as shown in Algorithm 2, the iterative algorithm is used to calculate the similarity. Then, one question is that how many iterations are required in our algorithm. Figure 3 shows the performance of our method on each website. Its horizontal axis represents the number of iterations and the vertical axis represents the F-score. And for each website, the F-scores with different α (range from 0.2 to 0.8) are also reported in this figure. From Figure 3, it is very easy to find that the best iteration number is 3. For almost all the websites, the performance of our method achieves the maximal values and converges after the third iteration. In addition, Figure 3 also indicates that our method is robust for different websites. In the following experiments, the iteration number is set to 3.

Next, let's turn to the discount factor α. Figure 4 reports the experimental results on the whole dataset. Here, the horizontal axis represents the discount factor α and the vertical axis represents the F-score. $\alpha = 0$ means that only the internal similarity is used in the algorithm, so the F-score equals to that in Figure 2 when $\beta = 0.6$. On the contrary, $\alpha = 1$ means that only the external similarity is used in the method, and the F-score is 80.20%. The performance is lower than the baseline system when only the external link information is used, but it is much better than the performance of the content-based method and structure-based method whose F-scores are 64.82% and 64.0% respectively. Besides, it is shown from Figure 4, the performance is improved significantly when the internal and external similarity measures are combined together. Furthermore, it is somewhat surprising that the discount factor α is not important as we previously expected. In fact, if we discard the cases that α equals to 0 or 1, the difference between the maximum and minimum F-score will be 0.76% which is very small. This finding indicates that the internal and external similarity can easily be combined and we don't need to make many efforts to tune this parameter when our method is applied to other websites. The reason of this phenomenon is that, no matter how much weight (i.e., 1- α) was assigned to the internal similarity, the internal similarity always provides a relatively good initial

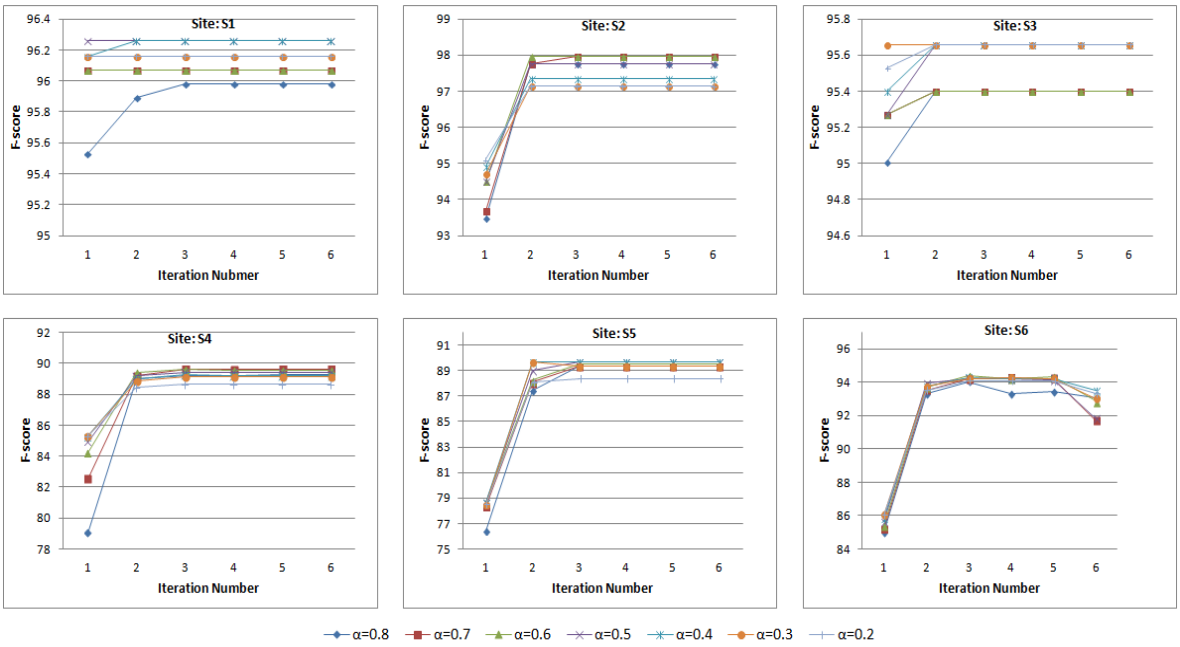

Figure 3 Experiment results of our method on each website

iterative direction. In the following experiments, the parameter α is set to 0.6.

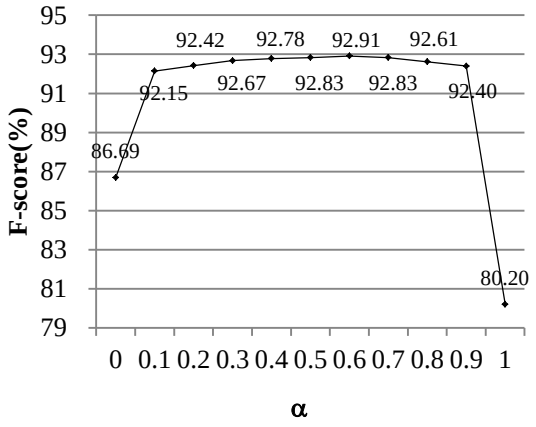

Figure 4 The F-scores of our method with different the value of α

The weight of pages

The weight of the neighbor pages should also be considered. For example, in the most websites, it is very common that most of the web pages contain a hyperlink which points to the homepage of the website. While in most of the English/Chinese websites, almost every English page will link to the English homepage and each Chinese page will point to Chinese homepage. The English and Chinese homepages are probably parallel, but they will be helpless to find parallel web pages, because they are neighbors of almost every page in the site. On the contrary, sometimes the parallel homepages have negative effects on finding parallel pages They will increase the translation similarity of two pages which are not indeed mutual translations. So it is necessary to amend the Algorithm 1.

The weight of each page is calculated according to its popularity:

$$w(p) = log\frac{N+c}{Freq(p)+c} \quad (6)$$

where $w(p)$ indicates the weight of page p, N is the number of all pages, $Freq(p)$ is the number of pages pointing to page p and c is a constant for smoothing.

In this paper, the weights of pages are used in two ways:

Weight 1: The 9th line of Algorithm 1 is amended by the page weight as follows:

$sum \leftarrow sum + ETS^{i-1}(x,y) \cdot (w(x)+w(y))/2$

Weight 2: The pages with low weight are removed from the input of Algorithm 1.

The experiment results are shown in Table 2.

Table 2 The effect of page weight

Type	No Weight	Weight 1	Weight 2
F-score (%)	92.91	92.78	92.75

Surprisingly, no big differences are found after the introduction of the page weight. The side effect of popular pages is not so large in our method. In the neighbor pages of a certain page, the popular pages are the minority. Besides, the iterative process makes our method more stable and robust.

The impact of the size of bilingual lexicon

The baseline system mainly combines the content-based similarity with structure similarity.

And two kinds of similarity measures are also used in our method. As Ma and Liberman (1999) pointed out, not all translators create translated pages that look like the original page which means that the structure similarity does not always work well. Compared to the structure similarity, the content-based is more reliable and has wider applicability. Furthermore, the bilingual lexicon is the only information that relates to the language pairs, and other features (such as structure and link information) are all language independent. So, it's important to investigate the effect of lexicon size in our method. We test the performance of our method with different size of the bilingual dictionary. The experiment results are shown in Figure 5. In this figure, the horizontal axis represents the bilingual lexicon size and the vertical axis represents the F-score. With the decline of the lexicon size, the performances of both the baseline method and our method are decreased. However, we can find that the descent rate of our method is smaller than that of the baseline. It indicates that our method does not need a big bilingual lexicon which is good news for the low-resource language pairs.

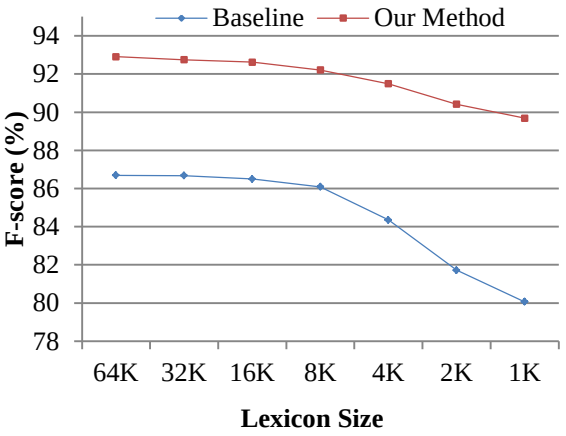

Figure 5 The impact of the size of bilingual lexicon

Error analysis

Errors occur when the two pages are similar in terms of structure, content and their neighbors. For example, Figure 6 illustrates a typical web page structure. There are 5 parts in the web page: U, L, M, R and B. Part M always contains the main content of this page. While part U, L, R and B always contain some hyperlinks such as "home" in part U and "About us" in part B. Links in L and R sometimes relate to the content of the page. For such a kind of non-parallel page pairs, let's assume that the two pages have the same structure (as shown in Figure 6). In addition, their content part M is very short and contains the same or related topics. As a result, the links in other 4 parts are likely to be similar. In this case, our method is likely to regard the two pages as parallel.

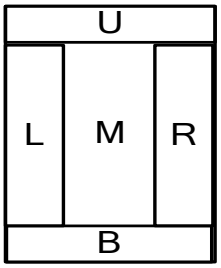

Figure 6 A typical web page structure

There are about 920 errors when our system obtains its best performance. By carefully investigating the error page pairs, we find that more than 90% errors fall into the category discussed above. The websites used in our experiments mainly come from Hong Kong government websites. Some government departments regularly publish quarterly or monthly work reports on one issue through their websites. These reports look very similar except the publish date and some data in them. The other 10% errors happen because of the particularity of the web pages, e.g. very short pages, broken pages and so on.

5 Conclusions and Future Work

Parallel corpora are valuable resources for a lot of NLP research problems and applications, such as MT and CLIR. This paper introduces an efficient and effective solution to bilingual language processing. We first explore how to extract parallel page pairs in bilingual websites with link information between web pages. Firstly, we hypothesize that the translation similarity of pages should be based on both internal and external translation similarity. Secondly, a novel iterative method is proposed to verify parallel page pairs. Experimental results show that our method is much more effective than the baseline system with 6.2% improvement on F-Score. Furthermore, our method has some significant contributions. For example, compared to previous work, our method does not depend on bilingual lexicons, and the parameters in our method have little effect on the final performance. These features improve the applicability of our method.

In the future work, we will study some method on extracting parallel resource from existing parallel page pairs, which are challenging tasks due to the diversity of page structures and styles. Besides, we will evaluate the effectiveness of our mined data on MT or other applications.

Acknowledgments

This research work has been sponsored by National Natural Science Foundation of China (Grants No.61373097 and No.61272259), one National Natural Science Foundation of Jiangsu Province (Grants No.BK2011282), one Major Project of College Natural Science Foundation of Jiangsu Province (Grants No.11KJA520003) and one National Science Foundation of Suzhou City (Grants No.SH201212).

The corresponding author of this paper, according to the meaning given to this role by School of computer science and technology at Soochow University, is Yu Hong

Reference

Chen, Jiang and Jianyun Nie. 2000. Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. Proceedings of the sixth conference on Applied Natural Language Processing, 21–28.

Resnik, Philip and Noah A. Smith. 2003. The Web as a Parallel Corpus. Meeting of the Association for Computational Linguistics 29(3). 349–380.

Kit, Chunyu and Jessica Yee Ha Ng. 2007. An Intelligent Web Agent to Mine Bilingual Parallel Pages via Automatic Discovery of URL Pairing Patterns. Web Intelligence and Intelligent Agent Technology Workshops, 526–529.

Zhang, Ying, Ke Wu, Jianfeng Gao and Phil Vines. 2006. Automatic Acquisition of Chinese-English Parallel Corpus from the Web. Joint Proceedings of the Association for Computational Linguistics and the International Conference on Computational Linguistics, 420–431.

Nie, Jianyun, Michel Simard, Pierre Isabelle and Richard Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 74–81.

Ma, Xiaoyi and Mark Y. Liberman. 1999. BITS: A Method for Bilingual Text Search over the Web. Machine Translation Summit VII.

Chen, Jisong, Rowena Chau and Chung-Hsing Yeh. 2004. Discovering Parallel Text from the World Wide Web. The Australasian Workshop on Data Mining and Web Intelligence, vol. 32, 157–161. Dunedin, New Zealand.

Jiang, Long, Shiquan Yang, Ming Zhou, Xiaohua Liu and Qingsheng Zhu. 2009. Mining Bilingual Data from the Web with Adaptively Learnt Patterns. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 2, 870–878.

Yanhui Feng, Yu Hong, Zhenxiang Yan, Jianmin Yao and Qiaoming Zhu. 2010. A novel method for bilingual web page acquisition from search engine web records. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 294–302.

Zhao, Bing and Stephan Vogel. 2002. Adaptive Parallel Sentences Mining from Web Bilingual News Collection. IEEE International Conference on Data Mining, 745–748.

Smith, Jason R., Chris Quirk and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 403–411.

Bharadwaj, Rohit G. and Vasudeva Varma. 2011. Language independent identification of parallel sentences using wikipedia. Proceedings of the 20th International Conference Companion on World Wide Web, 11–12. Hyderabad, India.

Gusfield, Dan. 1997. Algorithms on Strings, Trees and Sequences: Computerss Science and Computational Biology. Cambridge University Press

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