Journal of Computational Information Systems 9: 6 (2013) 2175–2182
Available at http://www.Jofcis.com

A Novel Method for Parallel Resources Acquisition from Bilingual Web Pages

Wei TANG*, Yu HONG, Yanhui FENG, Jianmin YAO, Qiaoming ZHU

School of Computer Science and Technology, Soochow University, Suzhou 215006, China

Abstract

A new approach is proposed for acquiring parallel resources by expanding seeds (some high-quality parallel sentence pairs) in the same page. Specifically, given a bilingual web page, the method is composed of three challenging tasks: 1) page's content structure is detected for adaptive viewing on its main data region; 2) some high-quality parallel sentence pairs in the main data region are selected as seeds for the next task; 3) wrappers for mining parallel resources are learnt by expanding the seeds. We identify and extract more parallel resources in the same page using the learnt wrappers. Our method can be applied to web documents written in any markup language and in any human language. The test is based on 500 bilingual web pages randomly selected from 12 popular English learning web sites, which gets a high F-score of 85.68%. The experimental results show that our method is quite promising.

Keywords: Parallel Resource; VIPS; SVM; Wrapper Construction

1 Introduction

Parallel resource are critical for many NLP applications, such as machine translation and cross language information retrieval. There have been extensive studies on parallel data extraction from parallel monolingual web pages of some bilingual web sites[1, 2]. 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, we observe that parallel resources may exist not only in two parallel monolingual web pages, but also in single bilingual web pages. People create such web pages for various reasons. For example, there are many English learning pages containing consistently formatted bilingual sentences (see Figure 1). The page* lists the most common oral sentences in English and their Chinese translation to facilitate English learning. To acquire bilingual web pages automatically, an effective method has been proposed in [3] by making use of search engines.

Based on these bilingual web pages, we use a seed-expansion method to mine interesting parallel resources based on the observation that parallel data usually appear collectively and follow similar formatting styles.

*http://english.51ielts.com/a/esp/office/2010/1029/30266.html

*Corresponding author.
Email address: sudajike@gmail.com (Wei TANG).

March 15, 2013


地道英语之办公室闲聊

2010-10-29 11:29:14 精品授权学校：新动力学校 浏览次数： [收藏此页]

当我们遇到办公室里的同事，就可以这样问：

Good morning! How are you today?

早上好！你今天好吗？

How are you this morning/afternoon/evening? (Being more specific may get a more specific answer)

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：

I\'m great actually—thanks for asking!

我很好，谢谢关心！

I\'m alright—a bit tired.

我还好，就是有点累。

Really good, thanks. I\'m getting excited for the weekend. (Then explain why you\'re excited)

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）

Can\'t complain!

好得没话说！

回答完后还可以问上一句：

How about yourself?

你如何呢？

Fig. 1: A snapshot of a bilingual web page

2 Related Work

2.1 Bilingual data mining from the web

As far as we know, most existing studies, such as [1], extract bilingual sentences from mined parallel documents using sentence alignment method. Munteanu[4] proposes a method to discover bilingual sentences in comparable corpora.

Cao[5] and Lin[6] propose methods to extract term translations based on the observation that translations of terms are often annotated in a pair of parentheses in many bilingual web pages, like “c1 c2 ... cn(e1 e2 ... em)”. However, the method may miss a lot of translations in the web.

Jiang[7] uses an adaptive pattern-based method to mine interesting parallel data based on the observation that parallel data usually appear collectively and follow similar surface patterns. Hong[8] attempts to directly search sentence pairs from the Web. We tackle the problem by formulating good search query using “Learning to Rank” and by filtering noisy document pairs.

2.2 Vision-based web page segmentation

Deng[9] proposes a top-down segmentation method called VIPS to extract page’s semantic structure based on its visual presentation. Such semantic structure is a hierarchical structure represented as a tree, in which each node corresponds to a block.

As to the high-level content block detection, Chen[10] proposes a method based on the HTML DOM tree, which gives the position and dimension information for each node in the DOM tree. Their high-level structure contains one or more of the following blocks: header, footer, left (right) side bar and body. Extracting the high-level content structure is to determine what extracted visual block falls into which high-level region. For instance, a side bar region depends on the width of the web page. So they define the left 1/4 part of a web page to be the left side bar region, and right 1/4 part to be the right side bar region.

2.3 Wrapper-based set expansion

Set expansion refers to expanding a given partial set of objects into a more complete set, which is closely related to the problem of unsupervised relation learning from the web[11]. Google SetsTM is a well-known example of a web-based set expansion system, which is a proprietary method that may be changed at any time. Richard[12] proposes a method called SEAL, which uses a novel technique to automatically construct wrappers that contains the seed sets. Such wrappers that bracket at least one occurrence of every seed on the page are learnt. Then they get all entity candidates from the extracted entity mentions by such wrappers.

3 Overview of the Proposed Method

Web documents are usually structured consistently within the same page but not across multiple pages. For one bilingual web page, we try to mine parallel resources within it by learnt wrappers. As illustrated in Figure 2, our system is comprised of three major components: the Page Analyzer, the Seed Extractor, and the Wrapper Constructor.

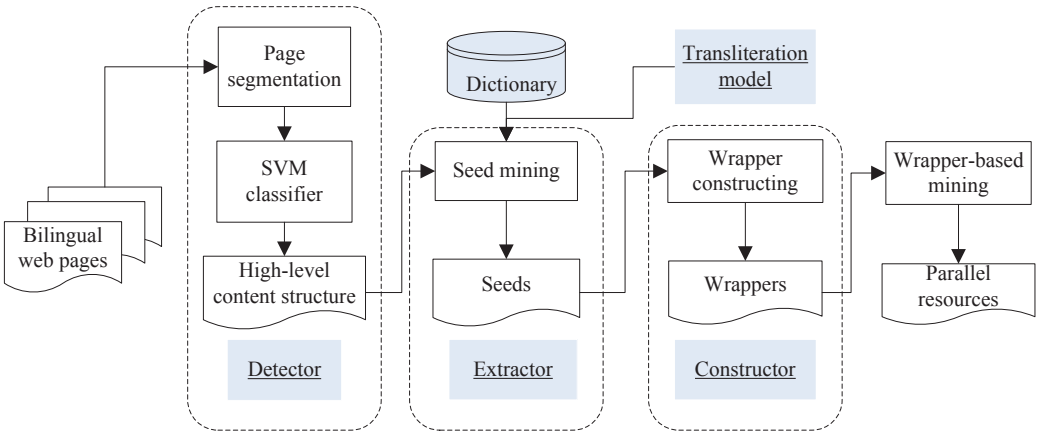

Fig. 2: Overview of the method

4 The Page Analyzer

4.1 Vision based page segmentation

VIPS algorithm utilizes the fact that semantically related contents are often grouped together, and the entire page is divided into different regions using visual separators such as images, lines, font sizes, blank areas, etc. VIPS iteratively uses DOM structure and visual cues for block extraction, separator detection and content structure generation.

In our method, VIPS algorithm is applied to a given web page to get an initial content structure tree. Considering each node of a tree as a unit semantic block, then we check the block belongs to which high-level region.

4.2 High-level structure detection

In this paper, the high-level structure of one page would contain one of the following two kind blocks: Body and Non-Body. Extracting the high-level content block is to determine what visual blocks falls into which high-level region. We try to detect the high-level structure of a web page by a SVM classifier, and all the features used are shown in Table 1.

In Table 1, Shape(B) is the shape of block based on height/width. RePos(B) is the relative position of the block's center point in the page. Per(IMG), Per(L), Per(LT), Per(T) and Per(ET) are relative feature functions about the inner images, hyper links, link texts, plain texts and the English texts in the given block. When creating a web page, Non-Body region usually has a lot of links or images and Body region usually locates in the most obvious part of a web page with main content. In this paper we only focus on blocks occupied by Body region, which can reduce the noise impacts. For easy reference, we will call the Body region as data region.

Table 1: All the features used in high-level structure detection

	Features	Remarks
1	$Shape(B) = \frac{B(h)}{B(w)}$	Block: B(h) and B(w) are the height and width of the block (B(c_x), B(c_y)) is the coordinate of block's center point. B(IMG) is the number of images in the block. B(L) is the number of links in the block. B(T) is the length of plain texts in the block. B(LT) is the length of link texts in the block. B(ET) is the length of English texts in the block. Page: P(h) and P(w) are the height and width of the page. P(IMG) are the number of images in the page. P(L) is the number of links in the page.
2	$RePos(B) = (\frac{B(c\_x)}{P(w)}, \frac{B(c\_y)}{P(h)})$	
3	$Per(IMG) = \frac{B(IMG)}{P(IMG)}$	
4	$Per(L) = \frac{B(L)}{P(L)}$	
5	$Per(LT) = \frac{B(LT)}{B(T)}$	
6	$Per(T) = \frac{B(T)}{P(T)}$	
7	$Per(ET) = \frac{B(ET)}{B(T)}$	

5 The Seed Extractor

5.1 Preprocessing

We firstly cut the input texts into continuous segments by language before extracting seeds, such as ⋯CECCCEECE⋯(C and E stand for Chinese and English terms respectively).

5.2 Seed extracting

The adjacent segment pairs are filtered by length-based trimming and word-overlap filtering. Pairs that meet the length-based threshold and with higher translation scores are selected as seeds for wrapper's construction.

(1) Length-based measure. In [13], the length measure is based on the assumption that each word in one language, L1, gives rise to a random number of words in the other language, L2. 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, s2, of this distribution, c is the expected number of characters in L2 per character in L1 and s2 is the variance of the number of characters in L2 per character in L1. They define:

$$\tau = \frac{l_2 - l_1 * c}{\sqrt{l_1 * s^2}} \tag{1}$$

We calculate c and s2 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. We testify 50 newly web pages and find that the precision and recall promising, so the length ratio calculator manner is robustness

(2) Word-Overlap measure. Word overlap judges the similarity of Chinese term and English term. The similarity Score(c_res,e_res) of Chinese term and English term is based on word-overlap as following:

$$Score(c\_res, e\_res) = \frac{\sum_{i=1}^{p} Max(Sim(c_i, e_j))}{\phi} \tag{2}$$

where the denominator is p+q, where p stands for the length of Chinese term and q is the length of English term. In addition, c_i stands for the ith word of Chinese term and e_j stands for the jth word of English term. $Sim(c_i, e_j)$ in [14] stands for the similarity of c_i and e_j. We rank all sentence pairs by Score(c_res, e_res) and choose top N pairs as valuable seeds.

6 The Wrapper Constructor

6.1 Algorithm

The information in semi-structured documents will be formatted quite differently on different pages, but fairly consistently within a single page. For each selected seed, let $s_{i,j}$ be the jth occurrence of ith seed. Let the left context $l_{i,j}$ and right context $r_{i,j}$ be the part of d preceding $s_{i,j}$ and following $s_{i,j}$. Let the middle context $m_{i,j}$ be the part of d between Chinese and English sentence. In this paper, $l_{i,j}$ and $r_{i,j}$ only consist of HTML tags, like <TagName>. One wrapper consists of three character strings, which specify the left, right and the middle context necessary for extracting parallel resources. For each pair of $l_{i,j}$ and $r_{i,j}$:

(1) **Context acquiring.** All possible suffixes of $l_{i,j}$ and all prefixes of $r_{i,j}$ are conceptually extracted; they are referred to as full suffixes and full prefixes respectively. We treat an HTML tag as a single unit when we extract all suffixes and all prefixes.

(2) **Candidate wrapper construction.** We simply concatenate any two elements in full suffixes and full prefixes as the left and right context of one newly candidate wrapper. All candidates have the same middle context $m_{i,j}$

(3) **Candidate wrapper ranking.** With the aim of cutting down the wrapper learning time, all the newly derived wrappers should be ordered by the number of tags in the left context, while some wrappers have the same left context they will be ordered by the number of tags in the right context.

(4) **Candidate wrapper filtering.** We deal with all the candidate wrappers by established order. If one candidate can bracket at least T pairs on the page, it will be selected as a desired wrapper, and other candidate wrappers constructed by the same seed will be removed.

Our approach is completely character-based and does not assume any language and domain. Also, we don't impose any limit on the length of the contextual strings nor do we require any parser.

6.2 Example

To simplify the wrapper construction process, suppose we select top 3 sentence pairs as seeds from the example page (in Figure 1), which are "Really good⋯excited) 非常好⋯兴奋）". Table 2 lists a context wrapper constructed by acquired contexts, with "[C]"and "[E]" representing the parallel sentence pairs. By this wrapper almost all parallel data in this page can be mined.

Table 2: Wrappers constructing from example page

Wrapper constructed from S1: $< divclass = "langs_en" > [E] < br/ >$ $< /div >< divclass = "langs_cn" > [C] < br/ >< /div >$
Extractions:
I'm alright-a bit tired. 我还好，就是有点累。
Really good, thanks. I'm getting excited for the weekend. (Then explain why you're excited)
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）
Can't complain! 好得没话说！

7 Experiments and Results

7.1 Evaluation on a human made test data set

We randomly select 500 bilingual web pages from 12 popular Chinese web sites and label all experimental data manually.

As mentioned before, we select top N seeds with higher translation score to construct wrappers, and as well as when selecting desired wrappers, we choose such candidates that bracket at least T pairs on the page. We aim to find a proper T, which is adaptive to different web pages. In order to extract more parallel resources, we set T to be 2. It is intuitive that the smaller T is,

the more the noises are. However, the products of wrappers are parallel sentence pairs with fixed middle context (acquired from seeds), which as well as are occupied in the data region. Based on the high-quality bilingual web pages collected by the method in [6], noises may be introduced with low possibility. The result shows that our F-score is acceptable. Compared to the method in [12], they train a classifier to select good patterns from candidates, but it is time and space consuming.

We study how the parameters N influence the performance of our method. The result is given in Table 3 based on their different settings. The model achieves satisfactory results when $N \geq 4$.

Table 3: Performance of different N

N	$P(\%)$	$R(\%)$	$F-score(\%)$
1	49.83	74.96	59.86
2	64.97	80.22	71.79
3	71.63	85.24	77.84
4	78.01	87.14	82.32
5	82.23	89.44	85.68
6	81.16	86.46	83.73
7	81.23	87.32	84.16

We select top N translation pairs as seeds for wrapper construction. As we know, the bigger the N is, the more the noises are. The results also demonstrate that our method can get better performance with higher value of N, which barely grows when N continuously increases. However, if all parallel sentences' translation score embedded by one wrapper are below N, we can not construct the wrapper successfully.As we know, the patterns learnt with the method of [12] are sensitive to the surface forms of sentence pairs, such as the punctuations in sentences and bullets in front of pairs. Their method also gets both a high precision and recall, and as well as its final F-score is 79.9% about exact mining. Our proposed method aims to mine parallel data in one page only by considering the layout similarity instead of taking the surface patterns into consideration, which is the major difference.

8 Conclusions

This paper presents a wrapper based method to acquire parallel resources automatically from bilingual web pages. The constructed wrappers are page, domain and language independent. A classifier is trained to detect the page's high-level content structure which narrow down the data region. We focus on the detected data region, which reduces noise (such as advertisements and navigations) effects on wrapper construction and wrapper-based mining. Experiments show that our algorithm has good performance.

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.

References

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

[2] Zhenxiang Yan, Yanhui Feng, Yu Hong, Jianmin Yao. Parallel Sentence Mining from the web. Journal of Computational Information Systems, 2009, 5: 6, pp. 1633 – 1641.

[3] 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.

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

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

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

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

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

[9] Cai, D., Yu, S. P., Wen, J. R., and W.-Y. Ma, Extrac- ting Content Structure for Web Pages Based on Visual Representation, Asia-Pacific Web Conference (2003), pp: 406 – 417.

[10] Y. Chen, W. Y. Ma and H.J. Zhang, "Detecting Web Page Structure for Adaptive Viewing on Small Form Factor Devices," Proc. 12th Int"l World Wide Web Conf. ACM Press, 2003, pp: 225 – 233.

[11] M. J. Cafarella, D. Downey, S. Soderland, and O. Etzioni, "KnowItNow: Fast, Scalable Information Extraction from the Web," in EMNLP, 2005.

[12] Richard C. Wang, William W. Cohen, Language Independent Set Expansion of Named Entities Using the Web, Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, pp: 342 – 350, October 28 – 31, 2007.

[13] William A. Gale, Kenneth W. Church, A program for aligning sentences in bilingual corpora, Com- putational Linguistics, v. 19 n. 1, March 1993.

[14] Dan Deng. 2004. Research on Chinese-English word alignment. Institute of Computing Technology Chinese Academy of Sciences, Master Thesis (in Chinese).