Journal of Computational Information Systems5:6(2009) 1633-1641
Available at http://www.JofCI.org

# Parallel Sentences Mining From The Web

Zhenxiang YAN, YanHui FENG, Yu HONG, Jianmin YAO†

*Provincial Key Laboratory of Computer Information Processing Technology, Soochow University, Suzhou 215006, China*

**Abstract**

Parallel sentences can benefit many NLP applications (e.g., machine translation, cross language information retrieval.) In this paper, the candidate bilingual webs pages are returned by submit sentence pairs to search engine and then validated by surface patterns. We propose an algorithm to candidate bilingual resource extraction and filter useless bilingual web pages. The pair sentences included in the candidate bilingual web pages is verified by a maximum entropy classifier combining length, word-overlap, alignment and text location features. Training sets and the mining seeds are acquired automatically. Experiment shows satisfactory parallel resource mining performance.

*Keywords:* Bilingual Web Page Selection; Bilingual Content Extraction; Parallel Sentence Verification

## 1. Introduction

Large-scale bilingual parallel corpus is very useful on some professional fields (e.g., paraphrase acquisition, OOV term translation, Machine Translation.)

**Related work**. In order to acquire more and more parallel sentences, many recent researches tried to build parallel corpora from bilingual web sites by making use of url strings or HTML tags as clues. Therefore, given a bilingual web site (e.g, www.gov.hk), (Ma and Liberman 1999, BITS; Chen and Nie 2000,PTMiner; Resnik and Smith 2003, STRAND; Zhang et al, 2006) utilized pre-defined url patterns to discover candidate parallel documents within the bilingual websites. Then content-based features and other features are used to verify the translation equivalence of the candidate pairs. (Lei Shi et al， 2006) mine the parallel web pages based on the HTML structure similarity of parallel web pages. (Long Jiang et al, 2009) mined bilingual data from the web with adaptively learnt patterns. Jian-Cheng Wu et al. (2005), Cao et al. (2007) and Lin et al. (2008) proposed different methods to extract term translations.

Just as the name implies, there is one thing for sure that the content of a Bilingual web page is written in two or more languages. Most of them have a primary language and a secondary language. The content in the secondary language is often the translations of some primary language text in the same page. Different from the previous methods, this paper tries to mine parallel sentences from these kind bilingual web pages

Mining and extracting parallel sentences is a very difficult task because of the nature of web documents. In our method the mining task is carried out step by step. The first step is deciding how to find and locate

† Corresponding author.
*Email addresses:* yzhenxiang@gmail.com (Zhenxiang YAN), fengyanhui456@gmail.com (YanHui FENG), hongy@suda.edu.cn (Yu HONG), jyao@suda.edu.cn (Jianmin YAO)


December, 2009

candidate bilingual web pages. The second step is proposing an effective filter filtering noisy content manner, because web pages often consist of many noises (e.g., advertisement, banner). Besides, we still need to do much relative work to verify whether a sentence pair is parallel or not. To improve working efficiency, we pay special attention to remove the useless bilingual web pages as much as possible.

This paper introduces how to mine parallel sentences from nature web pages in details. The rest of the paper is organized as follows. Section 2 introduces candidate bilingual web pages selection. In section 3 and section 4, we present how to extract candidate bilingual resources from the web pages and a useful algorithm to eliminate useless bilingual web pages. In section 5, we present parallel sentences extraction. Experiment and evaluation are presented in section 6.

## 2. Candidate Bilingual Web Page Selection

The purpose of candidate web pages selection is retrieval link of web pages contained parallel sentence pairs. We assume sentences pairs tend to co-occur in the same web pages. Hence querying a search engine with a pair of sentences, it is easy to obtain the candidate bilingual web page. We submit sentence pairs (i.e. E and C) to the search engine (e.g., http://www.baidu.com) and up to 1,000 web pages that contain the pair of sentences are returned. If a snippet search engine return contains two or more parallel sentence pairs, it's more likely to point bilingual web page.

**Example 1.** Sentence pairs "I see 我明白了" were sent to search engine, a snippet was returned as shown in Fig 1, among which, the first url does not point to bilingual pages, the rest are what we want. to detect whether the url point to the bilingual web pages, we define the surface patterns "…CE…CE…", if a snippet meets the surface pattern, the web page point by it is a candidate web pages. In the Fig 1, the first snippet surface pattern dose not accord with the defined pattern; so it isn't the candidate bilingual web page.


I see,疯狂英语每日开讲第9课-疯狂英语每日开讲-在线学英语口语
I see,疯狂英语每日开讲第9课疯狂英语每日开讲,最地道的美式英语口语在线学习网站,大量常用英语口语练习内容,彻底告别哑巴英语,迅速提高口语交际水平.
www.777cm.com/html/Reading/200904/3670.html 16K 2009-8-19 - 百度快照

100句最常用英语口语精典荐
1. i see. 我明白了。 2. i quit! 我不干了! 3. let go! 放手! 4. me too. 我也是。 5. my god! 天哪! 6. no way! 不行! 7. come on. 来吧(赶紧) 8. hold on. 等一
www.mingoouterwear.com.cn/20040227/155396 ... 8K 2008-4-25 - 百度快照
www.mingoouterwear.com.cn 上的更多结果

Fig.1 An Example of Snippet Search Engine Returns


奥运英语口语1000句:你在北京感觉怎么样

How are you enjoying Beijing?

你在北京感觉怎么样?

A Hello, Neil. How are you enjoying Beijing?

A 尼尔，你好。你在北京感觉怎么样?

Fig.2 A Bilingual Web Pages

## 3. Candidate Bilingual Resource Extraction

When the candidate web pages are obtained, further work has to be done to filter out the noise information and extract the candidate bilingual resources from the web pages.

### *3.1. Preprocessing*

This process finished by two steps: Dom parse and html segmentation.

Html segmentation: DOM-based segmentation approaches are widely used these years for title discovery (YunHua Hu et. al, 2005) and adaptive content delivery. Here the blocks are removed which contains only one kind of language resources firstly.

### *3.2. Main Content Extraction from the Candidate HTML Blocks.*

For candidate bilingual HTML block selection, we get the main content HTML blocks from the HTML pages, however, not all noises have been filtered. Our solution to this problem is to analyze the HTML structure of the candidate blocks.

1) Tag Path. Given a DOM tree of a web page, we fist select a tag path between root node and the node text node using Xpth, which is a language for addressing parts of a DOM document. Fig 2 shows a bilingual web page example. The path of English sentence "A Hello, Neil, How are you enjoying Beijing?" is "<html><head>….<table><tr> <p>" as the Chinese sentence "A 尼尔，你好。你在北京感觉怎么样？". From the example, the bilingual main content is in accordance with other main content in tag path. Thus, if a tag path of a text node appears more than 5 times, we think the content is candidate main content.

2) Locating in the identical text node. According to observation, if the Chinese sentence and the sentence are located under the same node, they are likely to be parallel. The inner text of these nodes will look like "….ECECEC…" Two adjacent snippets in different languages (indicated as "EC" and "CE") are considered a candidate parallel sentence pair.

3) Connect separate sections into an integrated sentence. Many authors assign various attributes (e.g., colors, fonts, styles) to key terms for emphasis. In the HTML pages, the key terms are under a new tag. That means a sentence will be divided into sections by the new tag node. So we should connect all the divided sections into an integrated target. Three kinds of examples are shown in figure 3.

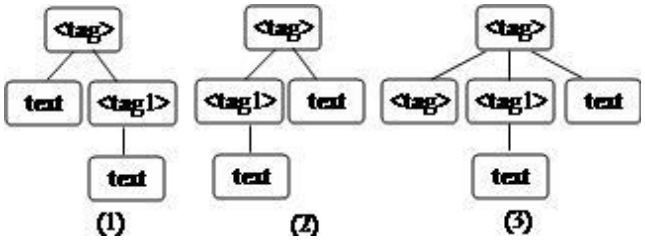


Fig.3 A Sentence is Divided into Several Sections by a New Tag Node

In the figure, for conditions (1) (2), to emphasize the term located under tag1 node, an integrated sentence is divided into two parts by tag1, In condition (3), tag1 divides the sentence or paragraph into three parts.

## 4. Eliminate Useless Bilingual Web Pages

The web has undergone at an exponential growth since its birth, and this expansion has generated a number of problems, for example, documents that are identical or almost identical are more and more in the internet. For example, some useful bilingual materials (e.g., classic spoken English translations) are reshipped by English learners. Thus, an effective algorithm to eliminate useless bilingual web pages is necessary.

We perform our deleting algorithm in two phases. In the first phage, we calculate a finger print value (M.O. Rabin, 1981) cValue to Chinese sentence and a finger value eValue to English sentence for a new parallel sentence pair respectively. Our mined parallel sentences were expressed in below style:

cValue 1 eValue 1
cValue 2 eValue 2

…. ….

cValue n eValue n

Cvalue n stands and Evalue n stands for Chinese and English sentences finger print value for the nth parallel sentence respectively. In the next phase, finger print values are calculated for sentences of a new bilingual web page. Calculate finger values for the whole sentences to stand for the bilingual web pages information is waste too long time. Useful sentences can be selected to stand for the bilingual web pages. In our work, we select at most 10 Chinese sentences and at most 20 English sentences randomly as test subset and calculate their finger print values; a bilingual web page are retained to do parallel sentences extraction, whose finger print values are in the mined parallel sentences collection mostly.

## 5. Parallel Sentences Extraction

This step is accomplished by preprocessing length-based and word-overlap filtering and maximum entropy classifier selecting.

### *5.1. Preprocessing*

1) Repair wrong abbreviation words: Repair some wrong abbreviation words, e.g., “Im” “arent.” “Ive”.
2) Delete noisy words. Such as “【例如】” “★”. 3) Amendment half-width to full-angle punctuation.

### *5.2. Length-based Measure*

The distance measure 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, $L_2$. lenScore(L1, L2) is defined below.

$$\text{len}Score(l1,l2) = \frac{length(l1)+1}{length(l2)+1}$$

The functions length(l) stands for the number of words of sentences. For the sake of setting the optimal parameters we test 1,500 pairs of parallel sentences, and the maximum and minimum values as threshold.

### *5.3. Word-overlap Score*

For a Chinese-English pair remained after the above length-based trimming, word-overlap (Liu Feifan, 2003; Deng Dan, 2004)can help us decide whether to align them or not. The similarity Score(c_res,e_res) of Chinese sentence and English sentence based on word-overlap as following.

$$Score(c\_res, e\_res) = (\sum_{i=1}^{p} \underset{1\le j\le q}{Max}(Sim(c_i, e_j)))/\phi$$

Where $\phi$ is normalization factor, it can be a: $p+q$ b: $p^2+q^2$ or c: $p$ stands for the length of Chinese sentence and $q$ stands for the length of English sentence. $c_i$ and $e_j$ stands for the ith word of Chinese word of Chinese sentence jth word of English word of English sentence respectively. $Sim(c_i, e_j)$ (Liu Feifan, 2003; Deng Dan, 2004) stands for the similarity of Chinese word $c_i$ and English word $e_j$.

The above filtering removes most of the noises that are not filtered by the candidate bilingual resource region selection. The dictionary does not contain all necessary entries, so good pairs may fail to pass the two filters. But those pairs cannot have been handled reliably anyway. So as a whole the overall effect of the filter is to improve the precision and robustness of the system. However, the filter also accepts many wrong pairs, because the word-overlap condition is weak. For instance, high frequency words almost always have a translation on the other side. An erroneous candidate sentence pair will be selected, if a few of the content words in this sentence happen to match the word-overlap threshold and the length-based trimming threshold. In our ways, parallel sentences verification based maximum entropy classifier involves general, alignment and location features solves this problem effectively.

### *5.4. Selecting the Best Parallel Sentences*

In bilingual web pages, a sentence or a paragraph may be located under several tags. As a result, all possible resource pairs those filtered by length-based trimming and word-overlap trimming are consist of full sentences and sentence fragments. So we need to select the best parallel sentences in two steps. Firstly, we connect sentence fragments into one full sentence, i.e. m Chinese sentence fragments form a full Chinese sentence with clear mean and N English sentence fragments constitute a complete English sentence. Sentence fragment usually doesn't have a stop symbol (e.g., full stop, question mark or an exclamation mark) and it isn't too long. Therefore, we check whether the sentence has stop symbols. If the sentence doesn't have stop punctuations, the next text node in the same language will be considered to connect to the former node. Secondly, we add the Chinese sentences and the English sentences into the possible resource pairs as the best parallel resources instead of former m Chinese sentence fragments and n English sentence fragments.

We focus on cases that m or n are lower than 3, and to calculate the similarity for candidate pairs using the formula below:

$$(\hat{m}, \hat{n}) = \text{argmax Score}(i...m, j...n) \quad m=0, 1, 2 \quad n=0, 1, 2$$

Where function Score(i...m,j...n) is the word overlap score between English resource (ith to mth text node) and Chinese resource (jth to nth text node). An example is listed below:

In table 1, the tag node (1) and tag node (2) constitute a complete English sentence. The best parallel sentence pair is not between the tag node (2) and (3), but between the complete English sentences and the tag node (3), as a result, the value of m is 2 in this example.

### *5.5. Parallel Sentences Verification*

For each candidate sentence pair, we need a reliable way to verify whether the pair is translation to each other. We treat this problem as a classification problem based on the word-overlap, word alignment and other features. The classification is carried out by a maximum-entropy classifier. For the classification problem, we have:

$$p(c_i \mid cp) = 1/Z(cp) \prod_{j=1}^{k} \lambda_j^{f_{ij}(c_i, cp)}$$

Where $c_i$ is the type of class ($c_0$ stands for "parallel", $c_1$ stands for "not parallel"). For the candidate pair *cp*, Z(cp) is a normalization factor, and $f_{ij}$ are the feature functions (indexed by class and feature). The

parameter values that maximize the likelihood of a given training corpus can be computed using various optimization algorithms (e.g, GIS).

Table 1 An Example of Selecting the Best Parallel Sentences

| |
|---|
| Ubuntu uses the Common UNIX Printing System ("CUPS") to handle printing. …….(1) |
| CUPS uses the Internet Printing Protocol ("IPP") as the basis for managing print jobs and queues. The Line Printer Daemon ("LPD") Server Message Block ("SMB"), and AppSocket (a.k.a. JetDirect) protocols are also supported with reduced functionality.  …….(2) |
| Ubuntu 使用通用 UNIX 打印系统（缩写“CUPS”）处理打印事务。CUPS 使用 Internet 打印协议（缩写“IPP”）作为管理打印作业和队列的基础。通过简化的功能也支持行式打印机服务、服务器消息块和 AppSocket（a.k.a. JetDirect）协议。 …….(3) |

**Location feature:** The location feature is an important factor. According to the discussion above, the Chinese sentence is often located nearby the English sentence. Two examples that the distance of the pair is 1 are listed and analyzed as below:

| | | | |
|---|---|---|---|
| (1) | <Segment1: Chinese1<br>Segment2: English1<br>Segment3:English2> | (2) | <Segment1: Chinese1<br>Segment2: English1<br>Segment3: Chinese2> |

In the first example, there are three sentence segments. Empirically, if Segment3 is in the same language with Segment2, the former pair <Segment1 and Segment2> is more likely to be parallel. In the second example, the three sentence segments are distributed in a different way. Segment3 is in a different language from Segment2.Compared to the pair<Segment1 and Segment2>, the similarity of the pair <Segment2 and Segment3> is higher, so this pair is more likely to be parallel.

To summarize, our classifier uses the following features to automatically compute the alignment between sentences.

Table 2 Features for the Maximum-entropy Classifier for Sentence Pair Verification

| Feature Type | Feature name | Description |
|---|---|---|
| General | Length ratio | The length measure of two resources (5.2) |
| | Word overlap | Word overlap score of two resource(5.3) |
| Alignment | Empty-Chinese | Percentage of Chinese words that have no connections |
| | Empty-English | Percentage of English words that have no connections |
| | Top fertility | The top three largest fertilities |
| Location | Location | The distance of two sentence |

In the table 2, fertility was defined in (Brown et al. 1993) as the number of words connected to a word in an alignment. (Dragos et al. 2005) think that words of high fertility are indicative of non-parallelism.

## 6. Experiment and Evaluation

**Acquire training set automatically:** The training instances for our classifier are 260 bilingual web pages labeled manually, which consist of 3,800 parallel sentence pairs after length-trimming and word-overlap filtering. Non-parallel sentence pairs are selected randomly from the Cartesian product of web language

resources. These non-parallel resource pairs those pass the above trimming and filtering process are taken as negative, and in our experiment 3,800 negative pairs are selected. To use the maximum entropy classifier effectively, the general feature, word alignment feature and distance feature are calculated automatically. The system get 19,000 parallel sentences, among which, the wrong pairs are deleted manually with 16,000 left as position training set. Finally, our system have total up to 20,000 position test sets and 20,000 negative pairs, which all pass the trimming and filtering process.

**Get the seeds used in candidate bilingual web pages automatically.** In our experiment, 1,000 parallel sentence pairs selected from our candidate bilingual web pages selection step are used as seeds and then are sent to search engines, and then we can extract parallel sentence pairs in the same way just as before. So far, 1590,000 sentences pairs have been obtained from 288,100 bilingual web pages in our loop iteration web mining ways. Our program will get more sentences pairs as our system runs. Manual evaluation of the system is performed as follows.

### *6.1. Evaluation on Overall Framework*

**Mining Accuracy**: The performance of the mining process is measured by precision and recall. 330 web pages are selected randomly from the output of our system, and two native Chinese speakers proficient in both Chinese and English are asked to evaluate the performance of our system. 3065 pair sentences mined from these web pages, the precision is 93% and the recall is 81%. The result shows that our system achieves high precision because the positive training examples and the testing instances are clean resources after the word-overlap filtering and the length trimming. Noisy pairs that do not pass the two filters are excluded. The low recall is mainly due to the word-overlap filter, because its performance relies on the dictionary, which doesn't contain all entries in the experiment sentences. However, the filter plays an important role in saving working time and improving precision.

**Mining Efficiency:** The seed used in current iteration are acquired from the result of the former iteration. Table 3 shows that, the iteration ways can get enough candidate bilingual pages, and as much as parallel sentences as it runs.

Table 3 Mining Efficiency

| Iteration | seeds | Candidate bilingual web page url | The number of mined parallel sentences |
|---|---|---|---|
| 1 | 1000 | 51,610 | 320,000 |
| 2 | 1000 | 41,260 | 26, 000 |
| 3 | 1000 | 56,000 | 35,000 |

### *6.2. Evaluation on the Contribution of the Different Features*

Table 4 Performance Changes with more Features Added to the Classifier

| Feature type | Precision | Recall |
|---|---|---|
| General features | 85% | 68% |
| General+ word alignment features | 91.5% | 74% |
| General+ word alignment +Location feature | 93% | 81% |

Table 4 shows that word alignment features and the location feature have significant effect on the mining performance. The dictionary in our experiment does not cover all entries and the high frequency words usually have translations on the other side. So the general features alone do not classify the resources well, which leads to 2 types of errors. One type of errors occur when the resources share many content words but express slightly different meanings, which lead to false positive result. The other occurs when the two sentences convey different amounts of information, and often contain partial translation in the other language. Word alignment specifies word mapping between the sentences and is helpful for parallel sentences verification. However, word alignment isn't enough accurate, so performance obtained by the general plus word alignment features is relatively lower than the one obtained by using three kinds of features. In bilingual web pages, Chinese sentences are located next to the English sentences in bilingual web pages. Thus, the location information has positive effects on our parallel resource selection.

## 7. Conclusions

Bilingual web pages have shown great potential as a source of up-to-data sentences which cover many domains and applications types. This paper has introduced a method for mining parallel sentences from the web. Firstly, based on the observation if snippets returned by search engine contain one or more parallel sentence pairs, i.e., it meets with our define surface patterns, therefore, candidate bilingual web pages url are got by submitting sentences pairs to the search engine and validated by our defined surface patterns. To help the system to save time, an effective finer print algorithm is used to eliminate useless bilingual web pages. Then, the main contents are extracted from the web pages segment based on the tag path, location information, link count, anchor text and resource integrity features. The length ratio and word-overlap information based on bilingual and synonym dictionaries are used to extract the potential parallel resources located under different tag nodes. At last the sentence pairs are verified by a maximum entropy classifier. In our work, the training sets and mining seed required by the system automatically .Experiment shows that word alignment and location features are key factors for precision and recall. The system gets a precision of 93% and extracts 1590,000 sentences pairs, as it running, it will get more sentence pairs.

## Acknowledgement

The work is supported by the National Natural Science Foundation of China under Grant No 60970057.

## References

[1] Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics

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

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

[4] Xiaoyi Ma and Mark Y.Liberman.1999.Bits:A Method for Bilingual Texts Search over the web. In Proceedings of Machine Translation SummitⅦ

[5] 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 Computational Linguistics and the International Conference on Computational Linguistics, Sydney, Australia

[6] Dekang Lin, Shaojun Zhao, Benjamin Van Durme, Marius Parsa,2008.Ming Parenthetical Translations from the web by Word Alignment In Joint Proceedings of the Association for Computational Linguistics and the International Conference on Computational Linguistics, Columbus, Ohio, USA, June 2008

[7] William and Keeneth W. Church.1993. A program for aligning sentences in Bilingual corpora. Computational Linguistics

[8] Peter F. Brown, Stephen A. I)ella Pietra, Vincent J. Della Pietra, and Robert L. Mercer.1993. The Mathematics of Statistical Machine Translation : Parameter Estimation Computational Linguisticss

[9] Yunhua Hu, Guomao Xin, Ruihua Song, Guop ing Hu, Shuming Shi, Yunbo Cao and Hang Li. Title Extraction from Bodies of HTML Documents and Its Application to Web Page Retrieval. SIGIR2005

[10] Chen Jiang and Jian-Yun Nie. 2000. Web parallel text mining for chineseenglish cross-language information retrieval.In International Conference onChinese Language Computing, Chicago, Illinois

[11] Jian-Cheng Wu, Tracy Lin, Jason S.Chang 2005.Learing Source-Target Surface Patterns for Web-based Terminology Translation ACL-2005

[12] Zhang, Y. and Vines, P. 2004. Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In the Proceedings of SIGIR2004

[13] W. Gale and K. Church. 1991. Identifying word correspondence in parallel text. In Proceedings of the DARPA NLP Workshop

[14] Feifan Liu , Jun Zhao, Bo Xu. 2003. Building Large-Scale Domain Independent Chinese-English Bilingual Corpus and the Researches on Sentence Alignment. JSCL

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

[16] Long Jiang, Shiquan Yang, Xiaohua Liu, Ming Zhou, Qingsheng Zhu. 2009. Mining Bilingual Data from the Web with Adaptively Learnt Patterns, ACL-IJCNLP 2009

[17] M.O. Rabin, Fingerprinting by random polynomials. Center for Research in Computing Technology,Harvard University, Report TR-15-81

[18] G.H. Cao, J.F. Gao and J.Y. Nie. 2007. A system to mine large-scale bilingual dictionaries from monolingualweb pages. MT summit. Pp: 57-64