

Mining Bilingual Data from the Web with Adaptively Learnt Patterns

Long Jiang¹, Shiquan Yang², Ming Zhou¹, Xiaohua Liu¹, Qingsheng Zhu²

¹Microsoft Research Asia
Beijing, 100190, P.R.China

²Chongqing University,
Chongqing, 400044, P.R.China

{longj, mingzhou, xiaoliu}@microsoft.com shiquany@gmail.com, qszhu@cqu.edu.cn

Abstract

Mining bilingual data (including bilingual sentences and terms¹) from the Web can benefit many NLP applications, such as machine translation and cross language information retrieval. In this paper, based on the observation that bilingual data in many web pages appear collectively following similar patterns, an adaptive pattern-based bilingual data mining method is proposed. Specifically, given a web page, the method contains four steps: 1) pre-processing: parse the web page into a DOM tree and segment the inner text of each node into snippets; 2) seed mining: identify potential translation pairs (seeds) using a word based alignment model which takes both translation and transliteration into consideration; 3) pattern learning: learn generalized patterns with the identified seeds; 4) pattern based mining: extract all bilingual data in the page using the learned patterns. Our experiments on Chinese web pages produced more than 7.5 million pairs of bilingual sentences and more than 5 million pairs of bilingual terms, both with over 80% accuracy.

1 Introduction

Bilingual data (including bilingual sentences and bilingual terms) are critical resources for building many applications, such as machine translation (Brown, 1993) and cross language information retrieval (Nie et al., 1999). However, most existing bilingual data sets are (i) not adequate for their intended uses, (ii) not up-to-date, (iii) apply only to limited domains. Because it's very hard and expensive to create a large scale bilin-

gual dataset with human effort, recently many researchers have turned to automatically mining them from the Web.

If the content of a web page is written in two languages, we call the page a Bilingual Web Page. Many such pages exist in non-English web sites. Most of them have a primary language (usually a non-English language) and a secondary language (usually English). The content in the secondary language is often the translation of some primary language text in the page.

Since bilingual web pages are very common in non-English web sites, mining bilingual data from them should be an important task. However, as far as we know, there is no publication available on mining bilingual sentences directly from bilingual web pages. Most existing methods for mining bilingual sentences from the Web, such as (Nie et al., 1999; Resnik and Smith, 2003; Shi et al., 2006), try to mine parallel web documents within bilingual web sites first and then extract bilingual sentences from mined parallel documents using sentence alignment methods.

As to mining term translations from bilingual web pages, Cao et al. (2007) and Lin et al. (2008) proposed two different methods to extract term translations based on the observation that authors of many bilingual web pages, especially those whose primary language is Chinese, Japanese or Korean, sometimes annotate terms with their English translations inside a pair of parentheses, like “ $c_1c_2\dots c_n(e_1 e_2 \dots e_m)$ ” ($c_1c_2\dots c_n$ is a primary language term and $e_1 e_2 \dots e_m$ is its English translation).

Actually, in addition to the parenthesis pattern, there is another interesting phenomenon that in many bilingual web pages bilingual data appear collectively and follow similar surface patterns. Figure 1 shows an excerpt of a page which introduces different kinds of dogs². The page provides

¹ In this paper terms refer to proper nouns, technical terms, movie names, and so on. And bilingual terms/sentences mean terms/sentences and their translations.

² <http://www.chinapet.net>

a list of dog names in both English and Chinese. Note that those bilingual names do not follow the parenthesis pattern. However, most of them are identically formatted as: “{Number}。{English name}{Chinese name}{EndOfLine}”. One exceptional pair (“1.Alaskan Malamute 啊拉斯加雪橇犬”) differs only slightly. Furthermore, there are also many pages containing consistently formatted bilingual sentences (see Figure 2). The page³ lists the (claimed) 200 most common oral sentences in English and their Chinese translations to facilitate English learning.

1.	Alaskan Malamute	啊拉斯加雪橇犬
2.	Beauceron	法国狼犬
3.	Bernese Mountain Dog	伯恩山地犬
4.	Bouvier des Flandres	比利时牧羊犬
5.	Boxer	拳师
6.	Bullmastiff	斗牛獒
7.	Cane Corso	卡斯罗
8.	Dobermann	杜宾
9.	Dogue de Bordeaux	波多尔
10.	Eskimo Dog	爱斯基摩犬

Figure 1. Consistently formatted term translation pairs


200句最实用的日常英语口语 从此不再做哑巴	
2006-02-05 08:39:00 百灵社区	
	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!

Figure 2. Consistently formatted sentence translation pairs

People create such web pages for various reasons. Some online stores list their products in two languages to make them understandable to foreigners. Some pages aim to help readers with foreign language learning. And in some pages where foreign names or technical terms are mentioned, the authors provide the translations for disambiguation. For easy reference, from now on we will call pages which contain many consistently formatted translation pairs Collective Bilingual Pages.

According to our estimation, at least tens of millions of collective bilingual pages exist in Chinese web sites. Most importantly, each such page usually contains a large amount of bilingual

data. This shows the great potential of bilingual data mining. However, the mining task is not straightforward, for the following reasons:

- 1) The patterns vary in different pages, so it's impossible to mine the translation pairs using predefined templates;
- 2) Some pages contain consistently formatted texts in two languages but they are not translation pairs;
- 3) Not all translations in a collective bilingual page necessarily follow an exactly consistent format. As shown in Figure 1, the ten translation pairs are supposed to follow the same pattern, however, due to typos, the pattern of the first pair is slightly different.

Because of these difficulties, simply using a classifier to extract translation pairs from adjacent bilingual texts in a collective bilingual page may not achieve satisfactory results. Therefore in this paper, we propose a pattern-based approach: learning patterns adaptively from collective bilingual pages instead of using the parenthesis pattern, then using the learned patterns to extract translation pairs from corresponding web pages. Specifically, our approach contains four steps:

- 1) Preprocessing: parse the web page into a DOM tree and segment the inner text of each node into snippets;
- 2) Seed mining: identify potential translation pairs (seeds) using an alignment model which takes both translation and transliteration into consideration;
- 3) Pattern learning: learn generalized patterns with the identified seeds;
- 4) Pattern based mining: extract all bilingual data in the page using the learnt patterns.

Let us take mining bilingual data from the text shown in Figure 1 as an example. Our method identifies “Boxer 拳师” and “Eskimo Dog 爱斯基摩犬” as two potential translation pairs based on a dictionary and a transliteration model (Step 2 above). Then we learn a generalized pattern that both pairs follow as “{BulletNumber}{Punctuation}{English term}{Chinese term}{EndOfLine}”, (Step 3 above). Finally, we apply it to match in the entire text and get all translation pairs following the pattern (Step 4 above).

The remainder of this paper is organized as follows. In Section 2, we list some related work. The overview of our mining approach is presented in Section 3. In Section 4, we give de-

³ <http://cul.beelink.com/20060205/2021119.shtml>

tailed introduction to each of the four modules in our mining approach. The experimental results are reported in Section 5 followed by our conclusion and some future work in Section 6.

Please note that in this paper we describe our method using example bilingual web pages in English and Chinese, however, the method can be applied to extract bilingual data from web pages written in any other pair of languages, such as Japanese and English, Korean and English etc.

2 Related Work

Mining Bilingual Data from the Web

As far as we know, there is no publication available on mining parallel sentences directly from bilingual web pages. Most existing methods of mining bilingual sentences from the Web, such as (Nie et al., 1999; Resnik and Smith, 2003; Shi et al., 2006), mine parallel web documents within bilingual web sites first and then extract bilingual sentences from mined parallel documents using sentence alignment methods. However, since the number of bilingual web sites is quite small, these methods can not yield a large number of bilingual sentences. (Shi et al., 2006), mined a total of 1,069,423 pairs of English-Chinese parallel sentences. In addition to mining from parallel documents, (Munteanu and Marcu, 2005) proposed a method for discovering bilingual sentences in comparable corpora.

As to the term translation extraction from bilingual web pages, (Cao et al., 2007) and (Lin et al., 2008) proposed two different methods utilizing the parenthesis pattern. The primary insight is that authors of many bilingual web pages, especially those whose primary language is Chinese, Japanese or Korean sometimes annotate terms with their English translations inside a pair of parentheses. Their methods are tested on a large set of web pages and achieve promising results. However, since not all translations in bilingual web pages follow the parenthesis pattern, these methods may miss a lot of translations appearing on the Web.

Apart from mining term translations directly from bilingual web pages, more approaches have been proposed to mine term translations from text snippets returned by a web search engine (Jiang et al., 2007; Zhang and Vines, 2004; Cheng et al., 2004; Huang et al., 2005). In their methods the source language term is usually given and the goal is to find the target language translations from the Web. To obtain web pages

containing the target translations, they submit the source term to the web search engine and collect returned snippets. Various techniques have been proposed to extract the target translations from the snippets. Though these methods achieve high accuracy, they are not suitable for compiling a large-scale bilingual dictionary for the following reasons: 1) they need a list of predefined source terms which is not easy to obtain; 2) the relevance ranking in web search engines is almost entirely orthogonal to the intent of finding the bilingual web pages containing the target translation, so many desired bilingual web pages may never be returned; 3) most such methods rely heavily on the frequency of the target translation in the collected snippets which makes mining low-frequency translations difficult.

Moreover, based on the assumption that anchor texts in different languages referring to the same web page are possibly translations of each other, (Lu et al., 2004) propose a novel approach to construct a multilingual lexicon by making use of web anchor texts and their linking structure. However, since only famous web pages may have inner links from other pages in multiple languages, the number of translations that can be obtained with this method is limited.

Pattern-based Relation Extraction

Pattern-based relation extraction has also been studied for years. For instance, (Hearst, 1992; Finkelstein-Landau and Morin, 1999) proposed an iterative pattern learning method for extracting semantic relationships between terms. (Brin, 1998) proposed a method called DIPRE (Dual Iterative Pattern Relation Expansion) to extract a relation of books (author, title) pairs from the Web. Since translation can be regarded as a kind of relation, those ideas can be leveraged for extracting translation pairs.

3 Overview of the Proposed Approach

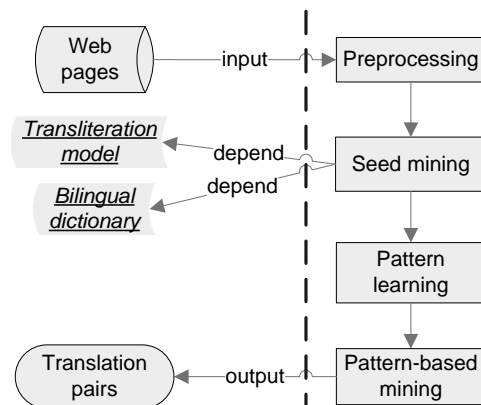


Figure 3. The framework of our approach

As illustrated in Figure 3, our mining system consists of four main steps: preprocessing, seed mining, pattern learning and pattern based mining. The input is a set of web documents and the output is mined bilingual data.

In the preprocessing step, the input web documents are parsed into DOM trees and the inner text of each tree node is segment into snippets. Then we select those tree nodes whose inner texts are likely to contain translation pairs collectively with a simple rule.

The seed mining module receives the inner text of each selected tree node and uses a word-based alignment model to identify potential translation pairs. The alignment model can handle both translation and transliteration in a unified framework.

The pattern learning module receives identified potential translation pairs from the seed mining as input, and then extracts generalized pattern candidates with the PAT tree algorithm. Then a SVM classifier is trained to select good patterns from all extracted pattern candidates.

In the pattern-based mining step, the selected patterns were used to match within the whole inner text to extract all translation pairs following the patterns.

4 Adaptive Pattern-based Bilingual Data Mining

In this section, we will present the details about the four steps in the proposed approach.

4.1 Preprocessing

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, where each node belongs to some predefined types (e.g. DIV, TABLE, TEXT, COMMENT, etc.). We removed nodes with types of “B”, “FONT”, “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.

Text Segmentation

After an HTML document is parsed, the inner text of each node in the DOM tree will be segmented into a list of text snippets according to their languages. That means each snippet will be labeled as either an English snippet (E) or a Chinese snippet (C).

The text segmentation was performed based on the Unicode values of characters⁴ first and then guided by the following rules to decide the boundary of a snippet under some special situations:

- 1) Open punctuations (such as ‘(‘) are padded into next snippet, and close punctuations (such as ‘)’) are padded into previous snippet; other punctuations (such as ‘;’) are padded into previous snippet;
- 2) English snippets which contains only 1 or 2 ASCII letters are merged with previous and next Chinese snippets (if exist). Since sometimes Chinese sentences or terms also contain some abbreviations in English.

Table 1 gives some examples of how the inner texts are segmented.

Inner text	China Development Bank (中国) 国家开发银行
Segmentation	China Development Bank (中国) 国家开发银行
Inner text	Windows XP 视窗操作系统 XP 版
Segmentation	Windows XP 视窗操作系统 XP 版

Table 1. Example segmentations (‘|’ indicates the separator between adjacent snippets)

Since a node’s inner text includes all inner texts of its children, the segmentation to all texts of a DOM tree has to be performed from the leaf nodes up to the root in order to avoid repetitive work. When segmenting a node’s inner text, we first segment the texts immediately dominated by this node and then combine those results with its children’s segmented inner texts in sequence.

As a result of the segmentation, the inner text of every node will look like “...ECECC⁵EC...”. Two adjacent snippets in different languages (indicated as “EC” or “CE”) are considered a Bilingual Snippet Pair (BSP).

Collective Nodes Selection

Since our goal is to mine bilingual knowledge from collective bilingual pages, we have to decide if a page is really a collective bilingual page. In this paper, the criterion is that a collective page must contain at least one Collective Node which is defined as a node whose inner text contains no fewer than 10 non-overlapping bilingual snippet pairs and which contains less than 10

⁴ For languages with the same character zone, other techniques are needed to segment the text.

⁵ Adjacent snippets in the same language only appear in the inner texts of some non-leaf nodes.

percent of other snippets which do not belong to any bilingual snippet pairs.

4.2 Seed Mining

The input of this module is a collective node whose inner text has been segmented into continuous text snippets, such as $\dots E_k C_h E_{k+1} C_{h+1} C_{h+2} \dots$. In this step, every adjacent snippet pair in different languages will be checked by an alignment model to see if it is a potential translation pair. The alignment model combines a translation and a transliteration model to compute the likelihood of a bilingual snippet pair being a translation pair. If it is, we call the snippet pair as a Translation Snippet Pair (TSP). If both of two adjacent pairs, e.g. $E_k C_h$ and $C_h E_{k+1}$, are considered as TSPs, the one with lower translation score will be regarded as a NON-TSP.

Before computing the likelihood of a bilingual snippet pair being a TSP, we preprocess it via the following steps:

- a) Isolating the English and Chinese contents from their contexts in the bilingual snippet pair. Here, we use a very simple rule: in the English snippet, we regard all characters within (and including) the first and the last English letter in the snippet as the English content; similarly, in the Chinese snippet we regard all characters within (and including) the first and the last Chinese character in the snippet as the Chinese content;
- b) Word segmentation of the Chinese content. Here, the Forward Maximum Matching algorithm (Chen and Liu, 1992) based on a dictionary is adopted;
- c) Stop words filtering. We compiled a small list of stop words manually (for example, “of”, “to”, “的”, etc.) and remove them from the English and Chinese content;
- d) Stemming of the English content. We use an in-house stemming tool to get the uninflected form of all English words.

After preprocessing, all English words form a collection $E = \{e_1, e_2, \dots, e_m\}$ and all Chinese words constitute a collection $C = \{c_1, c_2, \dots, c_n\}$, where e_i is an English word, and c_i is a Chinese word. We then use a linking algorithm which takes both translation and transliteration into consideration to link words across the two collections.

In our linking algorithm, there are three situations in which two words will be linked. The first is that the two words are considered translations of each other by the translation dictionary. The second is that the pronunciation similarity of the two words is above a certain threshold so that one can be considered the transliteration of the other. The third is that the two words are identical (this rule is especially designed for linking numbers or English abbreviations in Chinese snippets). The dictionary is an in-house dictionary and the transliteration model is adapted from (Jiang et al., 2007).

After the linking, a translation score over the English and Chinese content is computed by calculating the percentage of words which can be linked in the two collections. For some pairs, there are many conflicting links, for example, some words have multiple senses in the dictionary. Then we select the one with highest translation score.

For example, given the bilingual snippet pair of “Little Smoky River” and “小斯莫基河”, its English part is separated as “Little/Smoky/River”, and its Chinese part is separated as “小/斯/莫/基/河”. According to the dictionary, “Little” can be linked with “小”, and “River” can be linked with “河”. However, “Smoky” is translated as “冒烟的” in the dictionary which does not match any Chinese characters in the Chinese snippet. However the transliteration score (pronunciation similarity) between “Smoky” (IPA: s.m.o.k.i) and “斯/莫/基” (Pinyin: si mo ji) is higher than the threshold, so the English word “Smoky” can be linked to three Chinese characters “斯”, “莫” and “基”. The result is a translation score of 1.0 for the pair “Little Smoky River” and “小斯莫基河”.

4.3 Pattern Learning

The pattern learning module is critical for mining bilingual data from collective pages, because many translation pairs whose translation scores are not high enough may still be extracted by pattern based mining methods.

In previous modules, the inner texts of all nodes are segmented into continuous text snippets, and translation snippet pairs (TSP) are identified in all bilingual snippet pairs. Next, in the pattern learning module, those translation snippet pairs are used to find candidate patterns and then a SVM classifier is built to select the most useful patterns shared by most translation pairs in the whole text.

Candidate Pattern Extraction

First, as in the seed mining module, we isolate the English and Chinese contents from their contexts in a TSP and then replace the contents with two placeholders “[E]” and “[C]” respectively.

Second, we merge the two snippets of a TSP into a string and add a starting tag “[#]” and an ending tag “[#]” to its start and end. Following (Chang and Lui, 2001), all processed strings are used to build a PAT tree, and we then extract all substrings containing “E” and “C” as pattern candidates from the PAT tree. However, pattern candidates which start or end with “[E]” (or “[C]”) will be removed, since they cannot specify unambiguous boundaries when being matched in a string.

Web page authors commonly commit formatting errors when authoring the content into an html page, as shown in Figure 1. There, the ten bilingual terms should have been written in the same pattern, however, because of the mistaken use of “.” instead of “。”, the first translation pair follows a slightly different pattern. Some other typical errors may include varying length or types of white space, adjacent punctuation marks instead of one punctuation mark, and so on. To make the patterns robust enough to handle such variation, we generalized all pattern candidates through the following two steps:

- 1) Replace characters in a pattern with their classes. We define three classes of characters: Punctuation (P), Number (N), and White Space (S). Table 2 lists the three classes and the corresponding regular expressions in Microsoft .Net Framework⁶.
- 2) Merge identical adjacent classes.

Class	Corresponding regular expression
P	[p{P}]
N	[d]
S	[s]

Table 2. Character classes

For example, from the translation snippet pair of “7. Don’t worry.” and “别担心。”, we will learn the following pattern candidates:

- “[#][N][P][S][E][P][S][C][P]#”;
- “[N][P][S][E][P][S][C][P]#”;
- “[N][P][S][E][P][S][C][P]”;
- ...
- “[S][E][P][S][C][P]”;

⁶ In System.Text.RegularExpressions namespace

Pattern Selection

After all pattern candidates are extracted, a SVM classifier is used to select the good ones:

$$f_{\bar{w}}(\bar{x}) = \langle \bar{w}, \bar{x} \rangle$$

where, \bar{x} is the feature vector of a pattern candidate p_i , and \bar{w} is the vector of weights. $\langle \cdot, \cdot \rangle$ stands for an inner product. f is the decision function to decide which candidates are good.

In this SVM model, each pattern candidate p_i has the following four features:

- 1) **Generality**: the percentage of those bilingual snippet pairs which can match p_i in all bilingual snippet pairs. This feature measures if the pattern is a common pattern shared by many bilingual snippet pairs;
- 2) **Average translation score**: the average translation score of all bilingual snippet pairs which can match p_i . This feature helps decide if those pairs sharing the same pattern are really translations;
- 3) **Length**: the length of p_i . In general, longer patterns are more specific and can produce more accurate translations, however, they are likely to produce fewer matches;
- 4) **Irregularity**: the standard deviation of the numbers of noisy snippets. Here noisy snippets mean those snippets between any two adjacent translation pairs which can match p_i . If the irregularity of a pattern is low, we can be confident that pairs sharing this pattern have a reliably similar inner relationship with each other.

To estimate the weight vector, we extracted all pattern candidates from 300 bilingual web pages and asked 2 human annotators to label each of the candidates as positive or negative. The annotation took each of them about 20 hours. Then with the labeled training examples, we use SVM light⁷ to estimate the weights.

4.4 Pattern-based Mining

After good patterns are selected, every two adjacent snippets in different languages in the inner text will be merged as a target string. As we mentioned previously, we add a starting tag “[#]” and an ending tag “[#]” to the start and end of every target string. Then we attempt to match each of the selected patterns in each of the target strings and extract translation pairs. If the target

⁷ <http://svmlight.joachims.org/>

string was matched with more than one pattern, the matched string with highest translation score will be kept.

The matching process is actually quite simple, since we transform the learnt patterns into standard regular expressions and then make use of existing regular expression matching tools (e.g., Microsoft .Net Framework) to extract translation pairs.

However, to make our patterns more robust, when transforming the selected patterns into standard regular expressions, we allow each character class to match more than once. That means “[N]”, “[P]” and “[S]” will be transformed into “[\d]+”, “[\p{P}]+” and “[\s]+” respectively. And “[E]” and “[C]” will be transformed into “[^\u4e00-\u9fa5]+” (any character except Chinese character) and “.+”, respectively.

5 Experimental Results

In the following subsections, first, we will report the results of our bilingual data mining on a large set of Chinese web pages and compare them with previous work. Second, we will report some experimental results on a manually constructed test data set to analyze the impact of each part of our method.

5.1 Evaluation on a Large Set of Pages

With the proposed method, we performed bilingual data extraction on about 3.5 billion web pages crawled from Chinese web sites. Out of them, about 20 million were determined to contain bilingual collective nodes. From the inner texts of those nodes, we extracted 12,610,626 unique translation pairs. If we consider those pairs whose English parts contain more than 5 words as sentence translations and all others as term translations, we get 7,522,803 sentence translations and 5,087,823 term translations. We evaluated the quality of these mined translations by sampling 200 sentence translations and 200 term translations and presenting those to human judges, with a resulting precision of 83.5% for sentence translations and 80.5% for term translations.

As we mentioned in Section 2, (Shi et al., 2006) reported that in total they mined 1,069,423 pairs of English-Chinese parallel sentences from bilingual web sites. However, our method yields about 7.5 million pairs, about seven times as many.

We also re-implemented the extraction method using the parenthesis pattern proposed by (Lin et

al., 2008) and were able to mine 6,538,164 bilingual terms from the same web pages. A sample of 200 terms was submitted for human judgment, resulting in a precision of 78.5% which is a little lower than that of our original result. Further analysis showed that fewer than 20% of the bilingual terms mined with our method overlap with the data mined using the re-implemented method proposed by (Lin et al., 2008). This indicates that our method can find many translations which are not covered by the parenthesis pattern and therefore can be used together with the parenthesis pattern based method to build a bilingual lexicon.

Out of the term translations we mined, we found many which co-occur with their source terms only once in the Web. We check this by searching in Google with a Boolean query made of the term and its translation and then get the number of pages containing the query. If one attempts to extract this kind of low-frequency translation using a search engine-based method, the desired bilingual page which contains the target translation is not likely to be returned in the top n results when searching with the source term as the query. Even if the desired page is returned, the translation itself may be difficult to extract due to its low frequency.

5.2 Evaluation on a Human Made Test Data Set

Besides the evaluation of our method on a huge set of web pages, we also carried out some experiments on a human-constructed test data set. We randomly selected 500 collective nodes from the huge set of Chinese web pages and asked two annotators to label all bilingual data in their inner texts. Half of the labeled data are then used as the development data set and the rest as the test data set to evaluate our systems with different settings. Table 3 shows the evaluation results.

Setting	Type	Recall	Precision	F-Score
Without pattern	Exact	52.2	75.4	61.7
	Fuzzy	56.3	79.3	65.8
Without PG	Exact	69.2	78.6	73.6
	Fuzzy	74.3	82.9	78.4
With PG	Exact	79.3	80.5	79.9
	Fuzzy	86.7	87.9	87.3

Table 3. Performance of different settings

In Table 3, “Without pattern” means that we simply treat those seed pairs found by the alignment model as final bilingual data. “Without PG” and “With PG” mean not generalizing and generalizing the learnt patterns to class based form,

respectively. Evaluation type “Exact” means the mined bilingual data are considered correct only if they are exactly same as the data labeled by human, while “Fuzzy” means the mined bilingual data are considered correct if they contain the data labeled by the human.

As shown in Table 3, the system without pattern-based extraction yields only 52.2% recall. However, after adding pattern-based extraction, recall is improved sharply, to 69.2% for “Without PG” and to 79.3% for “With PG”. Most of the improvement comes from those translations which have very low translation scores and therefore are discarded by the seed mining module, however, most of them are found with the help of the learnt patterns.

From Table 3, we can also see that the system “With PG” outperforms “Without PG” in terms of both precision and recall. The reason may be that web writers often make mistakes when writing on web pages, such as punctuation misuse, punctuation loss, and extra spaces etc., so extracting with a strict surface pattern will often miss those translations which follow slightly different patterns.

To find out the reasons why some non-translation pairs are extracted, we checked 20 pairs which are not translations but extracted by the system. Out of them, 5 are caused by wrong segmentations. For example, “大提琴与小提琴双重协奏曲 Double Concerto for Violin and Cello D 大调第二交响曲 Symphony No.2 in D Major” is segmented into “大提琴与小提琴双重协奏曲”, “Double Concerto for Violin and Cello D”, “大调第二交响曲”, and “Symphony No.2 in D Major”. However, the ending letter ‘D’ of the second segment should have been padded into the third segment. For 9 pairs, the Chinese parts are explanative texts of corresponding English texts, but not translations. Because they contain the translations of the key words in the English text, our seed mining module failed to identify them as non-translation pairs. For 3 pairs, they follow the same pattern with some genuine translation pairs and therefore were extracted by the pattern based mining module. However, they are not translation pairs. For the other 3 pairs, the errors came from the pattern generalization.

To evaluate the contribution of each feature used in the pattern selection module, we eliminated one feature at a time in turn from the feature set to see how the performance changed in the absence of any single feature. The results are reported below.

Eliminated feature	F-Score (Exact)
Null	79.9
Generality	72.3
Avg. translation score	74.3
Length	77.5
Irregularity	76.6

Table 4. Contribution of every feature

From the table above, we can see that every feature contributes to the final performance and that Generality is the most useful feature among all four features.

6 Conclusions

Bilingual web pages have shown great potential as a source of up-to-date bilingual terms/sentences which cover many domains and application types. Based on the observation that many web pages contain bilingual data collections which follow a mostly consistent but possibly somewhat variable pattern, we propose a unified approach for mining bilingual sentences and terms from such pages. Our approach can adaptively learn translation patterns according to different formatting styles in various web pages and then use the learnt patterns to extract more bilingual data. The patterns are generalized to minimize the impact of format variation and typos. According to experimental results on a large set of web pages as well as on a manually made test data set, our method is quite promising.

In the future, we would like to integrate the text segmentation module with the seed mining and pattern learning module to improve the accuracy of text segmentation. We also want to evaluate the usefulness of our mined data for machine translation or other applications.

References

- P. F. Brown, S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19:2, 263-311.
- Sergey Brin. 1998. Extracting patterns and relations from the World Wide Web. In *Proc. of the 1998 International Workshop on the Web and Databases*. Pp: 172-183.
- 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.

- Chia-Hui Chang and Shao-Chen Lui. 2001. IEPAD: Inform extract based on pattern discovery. In Proc. of the 10th ACM WWW conference.
- Keh-Jiann Chen, Shing-Huan Liu. 1992. Word Identification for Mandarin Chinese Sentences. In the Proceedings of COLING 1992. Pp:101-107.
- Cheng, P., Teng, J., Chen, R., Wang, J., Lu, W., and Cheng, L. 2004. Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval. In the Proceedings of SIGIR 2004, pp 162-169.
- Michal Finkelstein-Landau, Emmanuel Morin. 1999. Extracting Semantic Relationships between Terms: Supervised vs. Unsupervised Methods. In Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure. Pp:71-80.
- Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In the Proceedings of COLING-92. Pp: 539-545.
- Huang, F., Zhang, Y., and Vogel, S. 2005. Mining Key phrase Translations from Web Corpora. In the Proceedings of HLT-EMNLP.
- L. Jiang, M. Zhou, L.-F. Chien, C. Niu. 2007. Named Entity Translation with Web Mining and Transliteration, Proceedings of the 20th IJCAI. Pp: 1629-1634.
- D. Lin, S. Zhao, B. Durme and M. Pasca. 2008. Mining Parenthetical Translations from the Web by Word Alignment. In ACL-08. pp 994-1002.
- Lu, W. and Lee, H. 2004. Anchor text mining for translation of Web queries: A transitive translation approach. ACM transactions on Information Systems, Vol.22, April 2004, pages 242-269.
- D. S. Munteanu, D. Marcu. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. 2005. Computational Linguistics. 31(4). Pp: 477-504.
- J-Y Nie, M. Simard, P. Isabelle, and R. Durand. 1999. Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of parallel Text from the Web. In SIGIR 1999. Pp: 74-81.
- Philip Resnik, Noah A. Smith. 2003. The Web as a Parallel Corpus. Computational Linguistics. 29(3). Pp: 349-380.
- Li Shao and Hwee Tou Ng. 2004. Mining new word translations from comparable corpora. In Proc. of COLING 2004. Pp: 618-624.
- Lei Shi, Cheng Niu, Ming Zhou, Jianfeng Gao. 2006. A DOM Tree Alignment Model for Mining Parallel Data from the Web. In ACL 2006.
- Jung H. Shin, Young S. Han and Key-Sun Choi. 1996. Bilingual knowledge acquisition from Korean-English parallel corpus using alignment method: Korean-English alignment at word and phrase level. In Proceedings of the 16th conference on Computational linguistics, Copenhagen, Denmark.
- J.C. Wu, T. Lin and J.S. Chang. 2005. Learning Source-Target Surface Patterns for Web-based Terminology Translation. ACL Interactive Poster and Demonstration Sessions,. Pp 37-40, Ann Arbor.
- Zhang, Y. and Vines, P.. 2004. Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In the Proceedings of SIGIR 2004. Pp: 162-169.