Mining Chinese-English Parallel Corpora from the Web

Bo Li
School of Computer Science
Wuhan University
Wuhan, 430072, China
whulibo@gmail.com

Juan Liu
School of Computer Science
Wuhan University
Wuhan, 430072, China
liujuan@whu.edu.cn

Abstract

Parallel corpora are a crucial resource in research fields such as cross-lingual information retrieval and statistical machine translation, but only a few parallel corpora with high quality are publicly available nowadays. In this paper, we try to solve the problem by developing a system that can automatically mine high quality parallel corpora from the World Wide Web. The system contains a three-step process. The system uses a web spider to crawl certain hosts at first. Then candidate parallel web page pairs are prepared from the downloaded page set. At last, each candidate pair is examined based on multiple standards. We develop novel strategies for the implementation of the system, which are then proved to be rather effective by the experiments towards a multilingual website.

1 Introduction

Parallel corpora consisting of text in parallel translation plays an important role in data-driven natural language processing technologies such as statistical machine translation (Brown et al., 1990) and cross-lingual information retrieval (Landauer and Littman, 1990; Oard, 1997). But the fact is that only a few parallel corpora with high quality are publicly available such as the United Nations proceedings and the Canadian Parliament proceedings (LDC, 1999). These corpora are usually small in size, specializing in narrow areas, usually with fees and licensing restrictions, or sometimes out-of-date. For language pairs such as Chinese and English, the lack of parallel corpora is more severe. The lack of such kind of resource has been an obstacle in the development of the data-driven natural language processing technologies. But the intense human labor involved in the development of parallel corpora will still make it very hard to change the current situation by hand.

The number of websites containing web pages in parallel translation increases considerably these years, which gives hope that we can construct parallel corpora with high quality in a big scale more easily. In this paper, we present a system named Parallel Corpus Mining System (PCMS) which can automatically collect Chinese-English parallel web corpora from the Web. Similar with previous work, PCMS uses a three-step process. First, the web spider WebZip[1] is used to crawl the hosts specified by users. In the second step, candidate parallel web page pairs are prepared from the raw web page set fetched based on some outer features of the web pages. A novel strategy is designed to utilize all these features to construct high quality candidate parallel page pairs, which can raise the performance and reduce the time complexity of the system. In the third step, candidate page pairs are evaluated based on multiple standards in which page structure and content are both considered. The actually parallel page pairs are saved.

The content-based strategy in the PCMS system is implemented mainly based on the vector space model (VSM). We design a novel implementation of VSM to bilingual text, which is called bilingual vector space model (BVSM). In previous content-based work, they usually use coarse criterions to measure the similarity of bilingual text. For exam-

[1] http://www.spidersoft.com/webzip/default.asp

ple, Ma and Liberman (1999) measured the content similarity by the count of parallel token pairs in the text which are weak at representing the actual content of the text. VSM was considered for evaluating the similarity of bilingual text in (Chen et al., 2004), but unfortunately the particular description of the implementation which was a bit complex was not mentioned in their work, and the time complexity of their system was rather high. Besides, there are also some other types of methods for mining parallel corpora from the web such as the work in (Resnik, 1998), (Resnik and Smith, 2003) and (Zhang et al., 2006). Most of these methods are unbalanced between precision and recall or computationally too complex. We detail the implementation of BVSM in the PCMS system in this paper. The experiments conducted to a specific website show that PCMS can achieve a better overall result than relative work reported.

The structure of the paper is as follows. The system architecture of PCMS is introduced in Section 2. We introduce the details of the step for preparing candidate web page pairs in Section 3. The next step, candidate page pair evaluation, is described in Section 4. We discuss the results of the experiments and conclude the paper in the last two sections.

2 The PCMS System

The PCMS system is designed to mine parallel corpora automatically from the web. As has been clarified above, the system employs a three-step process. The first is a web page fetching step. There are some tools to do the job and the PCMS system uses WebZip to fetch all the web pages from specific hosts. We usually choose some sites which probably contain high quality parallel web pages such as the site of the ministry of foreign affairs of China. After the web pages are obtained from the servers, the web pages which are too small, for example smaller than 5k bytes, are excluded from the page set. Then for each page in the page set, the HTML source of the web page is parsed and the noise such as the advertisement is excluded from the raw web page. The second is the candidate parallel page pair preparation step. The web pages are paired according to the URL similarity and some other features of the web pages. The third is the candidate parallel page pair evaluation step which is the key section of the PCMS system. Both web page structure and content are considered in this step. The candidate parallel page pairs prepared by the second step are first filtered by the structure-based criterion and then evaluated by the content-based criterion. We develop novel strategies for the third step and describe it in detail in the following sections.

3 Candidate Parallel Pair Preparation

The web spider can fetch a great many web pages in different languages from certain hosts. Usually the language of a web page can be identified by some feature strings of the URL. For example, the URLs of many English web pages contain strings such as *e, en, eng* and *english* which are called *language identification strings*. The *language identification strings* are usually attached to the other part of the URL with symbols such as '–', '/' and '_'. The number of web pages downloaded by the web spider is very large, so the pairs produced will be a huge amount if we treat each web page in language *A* and each in language *B* as a candidate pair, which will then make the third step of the system computationally infeasible. Parallel web pages usually have similar URLs. For example, the web page *P1* in Chinese and *P2* in English are parallel:

Web page *P1* URL[2]:
www.fmprc.gov.cn/chn/wjdt/wshd/t358904.htm
Web page *P2* URL:
www.fmprc.gov.cn/eng/wjdt/wshd/t358905.htm

We can see that the URL of page *P1* and the URL of page *P2* share most of the strings such as *www.fmprc.gov.cn*, *wjdt,* and *wshd*. In some other cases, the similarity between the URLs of parallel web pages may be not that direct but should still be obvious.

In PCMS, a novel strategy is designed to measure the URL similarity of the candidate web page pair. Before the URL similarity evaluation process, the *language identification strings* of the URLs should be substituted by a uniform string which seldom occurs in normal URLs. For example, the *language identification strings* such as *en, eng, cn* and *chn* are substituted by the string *** which seldom occurs in normal URLs. For example, the above page *P1* after the URL substitution process is *www.fmprc.gov.cn/***/wjdt/wshd/t358904.htm*. After the substitution process, the similarity of the

[2] The protocol string HTTP is omitted here.

new URLs is evaluated. For evaluating the URL similarity of web page *PA* in language *A* and web page *PB* in language *B,* the following criterions are considered.

Criterion 1: *URL length difference.*

It can be found that the length of the URLs of parallel web pages is usually similar. The length of the URL here refers to the number of directories in the URL string. For example, the URL of the above web page *P1* contains the directories ***[3], *wjdt* and *wshd*, and then the URL length of *P1* is 3. If two web pages *PA* and *PB* are parallel, the URL length of *PA* and *PB* should be similar. The *URL length difference* criterion is define as

$$URL\ diff\ (PA,PB)=\frac{|len(PA)-len(PB)|}{len(PA)+len(PB)} \quad (1)$$

where *URL diff(PA,PB)* is the *URL length difference* between *PA* and *PB, len(PA)* is the URL length of page *PA* and *len(PB)* is the URL length of *PB*. The value of *URL length difference* is between 0 and 1, and the more similar two URLs are, the smaller the value is. If the URL lengths of *PA* and *PB* are the same, the *URL length difference* between *PA* and *PB* should be 0.

Criterion 2: *URL directory similarity.*

Besides URL length, URL directory information is also considered in the candidate page pair preparation step. It can be observed that the URLs of parallel web pages usually share similar directory structure which can be represented by the common directories in the URLs. For example, the above web page *P1* and web page *P2* share the directories ****, wjdt* and *wshd.* To measure the *URL directory similarity* of the web page *PA* and the web page *PB*, a criterion is defined as

$$URL\ dirsim(PA,PB)=\frac{2*comdir(PA,PB)}{len(PA)+len(PB)} \quad (2)$$

where *URL dirsim(PA,PB)* is the *URL directory similarity* of page *PA* and page *PB*, *comdir(PA,PB)* is the number of common directories *PA* and *PB* share, *len(PA)* and *len(PB)* are the same as above. The value of *URL directory similarity* is between 0 and 1. The bigger the value is, the more similar the two pages are. When two web pages have the same URLs, the *URL directory similarity* should be 1.

[3] The *language identification strings* of the URL have been substituted by the uniform string ***.

Criterion 3: *Similarity of some other features.*

Some other features such as the file size of the web page and the time the page created can help to filter the nonparallel web page pairs with low cost.

Based on the combination of the above criterions, the web page pairs of which the similarity excesses certain threshold are treated as the candidate parallel pairs, which are then to be processed by the following evaluation step.

4 Candidate Parallel Pair Evaluation

It is the key section of the system to evaluate the candidate parallel web page pairs. Though content-based methods are what the candidate parallel page pair evaluation step mainly relies on, the structure of the web pages is also considered in the evaluation step of the PCMS system for it can help to filter out some page pairs that are obviously non-parallel at low cost. The candidate parallel page pair set is first filtered by the structure-based strategy which is similar with the one in (Resnik, 1998), and we consider some more structure relative features such as color and font. A loose constrain is set on the structure similarity criterion, because it is merely a preliminary filter step to reduce the scale of the problem.

After the structure-based filter stage, the page pairs left are then to be evaluated by the content-based stage which is the key of the candidate parallel page pair evaluation step. The performance of the PCMS system relies mainly on this module. In the content-based stage, the candidate page pairs are first filtered based on some content related features and then the page pairs left are evaluated by the BVSM model.

4.1 The Content Related Feature-based Filter

In the first part of the content-based strategy, some content related features such as time stamp and navigation text are combined to construct a preliminary step to filter the candidate page pair set and reduce the number of pairs to be processed by BVSM. Many web pages contain time stamps which identify the time when the web pages were constructed. If two pages are parallel, the time when they are constructed should be similar. Navigation text usually demonstrates the type information of the content of the web page. For example, a web page with anchor text *Home-News-China* is probable about the news which happened in China.

So if two web pages are parallel, their navigation text if there is any should be similar. To evaluate the similarity of two pieces of navigation text in two languages, we need a bilingual navigation text wordlist. For each layer, for example *news*, in one navigation text, if its translation *新闻 xin-wen* appears in the other navigation text, the similarity *count* will be added by 1. The similarity between two pieces of navigation text is defined as

$$similarity = \frac{2 * count}{layer_{NC} + layer_{NE}} \quad (3)$$

where $layer_{NC}$ demonstrates the layer count of the navigation text of the Chinese web page and $layer_{NE}$ is that of the English web page. For example, the $layer_{NE}$ of the navigation text *Home-News-China* is 3. If the similarity gotten from formula (3) is below certain threshold, the corresponding web page pair will not be considered as parallel.

4.2 The BVSM Model

In the second part of the content-based strategy, BVSM is implemented to evaluate the similarity of candidate parallel page pairs. VSM is an important technology for representing text and has been applied to some other research areas. But this model is usually applicable to monolingual text processing problem. For bilingual text processing, we should design a new strategy to use VSM for the new problem. A bilingual dictionary is a must for importing VSM to bilingual problem. We give a brief introduction to the bilingual dictionary we use first. Each *entry line* of the dictionary consists of three parts. The first part is the English word, the middle is a list separator and the last is the corresponding Chinese word. A sample of the dictionary can be found in *Appendix A*. For each English word, there may be some Chinese words serving as its translations. The same conclusion can be gotten for each Chinese word.

Based on the bilingual dictionary, we can represent the Chinese and English web pages as vectors respectively. First, we give every English word in the bilingual dictionary a unique ID according to its position in the dictionary beginning from 1. For example, the ID of the English word in the first row is 1, and the ID of the next new English word in the dictionary is 2 and so forth. For convenience, we denote the Chinese web page as *C* and the English web page as *E* in each web page pair. We then can represent each web page as follows.

For *E,* we extract all the words from the web page and stem them first. The length of the vector of *E* equals the length of the bilingual dictionary which is the number of the different English words in the dictionary. For each dimension of the vector, for example *k,* we assign the number of the words with ID *k* occurring in all the words extracted to it. If certain words in the bilingual dictionary never occur in *E,* we assign the value 0 to the corresponding dimensions which are identified by the IDs of those words. If some words in *E* haven't occurred in the dictionary, we just ignore them.

For *C*, the procedure to construct a vector is more complex. In the PCMS system, the procedures of word segmentation and POS for Chinese are finished in a single run. The length of the vector of *C* equals to that of the vector of *E*. As has been pointed out, one Chinese word may correspond to more than one English word in the bilingual dictionary. For example in *Appendix A,* the Chinese word *放弃 fang-qi* corresponds to *abandon*, *depart* and *leave*. In the vector of *E*, each dimension strands for the count of a single English word with a unique ID occurring in the English text. In order to construct a vector for *C* which is comparable to the vector of *E*, a single Chinese word in *C* should contribute to more than one dimension of the vector of *C*. In order to distribute the count/weight of each Chinese word to the corresponding dimensions of the vector of *C,* we first count the number of each entry which is a Chinese word with a specific POS, for example *(放弃, Verb)*, in *C*. Then for each entry, we distribute its count to all the dimensions identified by the IDs of the English words which the Chinese word in the entry corresponds to. The count distribution process is detailed below.

If the Chinese word in the entry *Cent* is a content word which we call here to mean that it carries the main content of a language including noun, verb and adjective, we will divide the corresponding English words in the bilingual dictionary into four separate classes: the words that haven't appeared in the English text (C_4), the words that have the same POS with the entry (C_1), the words that have similar POS with the entry (C_2) and the other words (C_3). For convenience, the count of the entry *Cent* in *C* is denoted as N_{1234}. If the capacity of C_4 is 0 which means there are no words belonging to the class C_4, then N_{1234} is all devoted to the words

in C_1, C_2 and C_3, else a certain proportion, for example 10%, of N_{1234} is assigned to all the words in C_4 averagely and the left of N_{1234} is assigned to the words in C_1, C_2 and C_3. Similarly, we denote the count left to words in C_1, C_2 and C_3 as N_{123}, and then if the capacity of C_3 is 0, N_{123} is all denoted to the words in C_1 and C_2, else a certain proportion of N_{123} is denoted to all the words in C_3 averagely and the left of N_{123} is devoted to the words in C_1 and C_2. For words in C_1 and C_2, the count distribution strategy is similar.

If the Chinese word in the entry *Cent* is not a content word, we classify the corresponding English words into two classes: the words that haven't appeared in the English text (C_2) and the other words (C_1). The same method as above is used to distribute the count.

4.3 Similarity Evaluation Criterions

Based on the above strategies, the two web pages can be represented by their vectors respectively. Then the next step is to calculate the similarity of the two vectors, which is also the similarity of the two web pages. Some comments were given on different similarity measures such as *Euclidence distance*, *Inner product*, *Cosine coefficient*, *Dice coefficient* and *Jaccard coefficient* in (Chen et al., 2004). It was suggested that for a pair of documents to be considered parallel, we could expect that these two documents contained the two corresponding sets of translated terms and each corresponding term was carrying an identical contextual significance in each of the document respectively. For that, the *Jaccard coefficient* is more appropriate for the calculation of the similarity score. While in our experiments, we find that *Cosine coefficient* is more suitable. Because the size of the bilingual dictionary is small and we exclude all the words which are not in the dictionary from the text of the web pages when we construct the vectors, it is possible that the counterparts of some words in one web page can not be found in its corresponding web page. Though we have done some smooth work in the BVSM model, there is still a gap between the assumptions by Chen et al. (2004) and the situation of our problem. The second reason we think is that the translation process by human is almost sentence to sentence, but not word to word. As a result, it is normal that there are no words in one language serving as the translation for certain words in the other language. Based on the *Cosine coefficient* criterion, the similarity between two vectors which are represented by $(x_1, x_2, x_3, \ldots, x_p)$ and $(y_1, y_2, y_3, \ldots, y_p)$ respectively is

$$\cos ine\, coefficient = \frac{\sum_{i=1}^{p} x_i y_i}{\sqrt{\sum_{i=1}^{p} x_i^2 * \sum_{i=1}^{p} y_i^2}} \qquad (4)$$

The similarity measure is between 0 and 1, and the bigger the value is, the more similar the two vectors are. We set a certain threshold for the similarity measure based on our experience in PCMS.

5 Experiments and Discussion

In this section, we practice the experiments designed to evaluate the performance of the PCMS system and compare it with similar work earlier.

5.1 Evaluation Standards

Precision and *recall* are two widely used evaluation standards in the area of natural language processing. In our experiments, we define *precision* as the proportion of page pairs in parallel translation to the total page pairs produced by the PCMS system. *Recall* is defined as the proportion of page pairs in parallel translation produced by the PCMS system to the total parallel page pairs in the whole web page set.

The number of pairs in parallel translation should be calculated from the human annotated page pairs. We ask a native Chinese speaker who has a fluent English tongue to annotate these page pairs. To calculate the *recall*, we need to know the number of parallel pairs in the web page set. It is hard to count out the actual number of the parallel pairs in the page set because the web page set is really too big. We build a relatively smaller test set to test the *recall* of the PCMS system.

5.2 Parallel Corpus Construction

In order to construct a high quality parallel corpus in the experiments, the website of the ministry of foreign affairs of China (*http://www.fmprc.gov.cn*) is chosen to be crawled. After the rough observation, it is found that a huge number of web pages fetched are in parallel translation. We get a web page set consisting of 40262 Chinese web pages and 17324 English web pages by the tool WebZip. After the preprocess step, the web pages left are to

be examined by the core modules of PCMS. It takes nearly 3 hours to finish the task on a PC with a P4 2.0G CPU and 512MB RAM, which is faster than the early systems. To evaluate the *precision* of the system, we randomly choose a subset of the web page pairs which PCMS gives as output, and get a web page set of 500 web page pairs. We manually annotate it and find that there are 479 truly parallel page pairs among them. Then the *precision* is about 96%. We analysis the 21 non-parallel pairs the PCMS system gives and find that most of these web pages are short web pages containing limited text. To obtain the *recall* of the PCMS system, we construct a test page set consisting of 350 parallel page pairs and 150 nonparallel page pairs. The ratio 350/150 is decided based on rough estimation of the whole page set. The PCMS system is examined on the test set, which produces 337 page pairs which are truly parallel, thus a *recall* of 96%. We analysis the 13 parallel pages which are recognized as nonparallel by the PCMS system and find that most of them are short web pages. We then come to the conclusion that the drawback that BVSM is weak at representing short text leads to the system's failure to identify the parallel web page pairs. Though the model has some drawbacks, the overall result consisting of performance and time complexity is much better than the former similar work.

6 Conclusion

The paper presents a web-based parallel corpus construction system PCMS. The system first fetches all the web pages from specific hosts, and then prepares candidate parallel web page pairs based on features such as URL and web page file size. At last the candidate pairs are examined by a two-stage similarity evaluation process in which the structure and content of the web pages are both considered. To enhance the performance of the PCMS system, we design some novel strategies for the implementation of these steps. The results of the experiments show the high performance and low time complexity of the PCMS system. All in all, the PCMS system is a reliable and effective tool for mining parallel corpora from the web.

References

Brown, P. F., Cocke, J., Pietra, S. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., et al. (1990). A statistical approach to machine translation. *Computational Linguistics, 16(2),* 79-85.

Chen, J., Chau, R., and Yeh, C. H. (2004). Discovering parallel text from the World Wide Web. In *Proc. of DMWI-04,* Dunedin, New Zealand.

Landauer, T. K. and Littman, M. L. (1990). Fully automatic cross-language document retrieval using latent semantic indexing. In *Proc. of the 6th Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research,* Waterloo, Ontario.

LDC. (1999). Linguistic Data Consortium (LDC) home page. http://www.ldc.upenn.edu

Ma, X. and Liberman, M. Y. (1999). BITS: A method for bilingual text search over the web. In *Proc. of the Machine Translation Summit VII.*

Oard, D. W. (1997). Cross-language text retrieval research in the USA. In *Proc. of the 3rd ERCIM DELOS Workshop,* Zurich, Switzerland.

Resnik, P. (1998). Parallel strands: A preliminary investigation into mining the web for bilingual text. In *Proc. of AMTA-98,* Langhorne, PA.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics, 29(3),* 349-380.

Zhang, Y., Wu, K., Gao, J., and Vines, P. (2006). Automatic acquisition of Chinese-English parallel corpus from the web. *In Proceedings of ECIR-06,* London.

Appendix A: A Sample Bilingual Dictionary

abandon --- 背弃
abandon --- 丢弃
abandon --- 放弃
abandon --- 抛弃
abc --- 初步
abc --- 入门
abc --- 字母
abc --- 基本
……
depart --- 出发
depart --- 放弃
depart --- 离开
depart --- 起程
……
leave --- 放弃
leave --- 离开
leave --- 离去
leave --- 留下
……