Collected Data Bilingual Automatic Vietnamese - Chinese From Websites

Minh Trinh[1], Quang Le[2], Phuoc Tran[3], Huy Pham[4]

[1,2,3]NLP-KD Lab, Faculty of Information Technology,

Ton Duc Thang University, Ho Chi Minh city, Vietnam

[1]Email:51403301@student.tdt.edu.vn

[2]Email:51403014@student.tdt.edu.vn

[3]Email: tranthanhphuoc@tdt.edu.vn

[4]Faculty of Information Technology,

Ton Duc Thang University, Ho Chi Minh city, Vietnam

Email:phamvanhuy@tdt.edu.vn

Abstract. Data monolingual-bilingual is extremely necessary for natural language processing, especially for machine translation. Chinese-Vietnamese is a resource-poor language a pair, with bilingual Chinese-Vietnamese data limited. Therefore, in this paper, we propose a method to automatically collect bilingual Chinese-Vietnamese documents from bilingual Chinese-Vietnamese websites. This bilingual textual material is the premise for extracting bilingual pairs in next research works. Our collection system was conducted on 10 bilingual sites of Vietnamese and Chinese and initially gave encouraging results. This system can be deployed to collect automatically for many other pairs of languages.

Keywords: Bilingual website, Chinese-Vietnamese texts, Natural Language processing, Collecting bilingual texts.

.

1 Introduction

Bilingual data warehouses are mandatory for modern machine translation approaches such as statistical machine translation or neural machine translation. All methods share the basic underlying principle of applying a translation model to capture the lexical translations and taking a language model to quantify fluency of the target sentence. Therefrom shows that the input data for statistic machine translation is very important and the quality of machine translation also depends on the data. Therefore, constructing a training corpus which contains a set of parallel sentences becomes one of the most important tasks in building any machine translation system.

To obtain data for machine translation, linguists often use one in two methods, manual collection or automatic text extraction from websites. The manual method is highly accurate but requires a team of linguistic and bilingual specialists. This manual method takes a lot of time and cost.

[3] Phuoc Tran is a corresponding author

Nowadays, with the development of bilingual websites, the method of automatic extraction is also highly effective. However, the accuracy level is not as high as the manual method but can also be used in the study. With this automated method will save time and costs and the number of words will be great.

In this paper, our purpose is to initially collect Vietnamese-Chinese bilingual data from bilingual web pages of the Vietnamese. We choose to collect data from our web pages because these pages are of various types such as sports, politics, culture, education, economy, life. Words and content of the pages of this magazine are plentiful, ranging from luxury to fork, so that the articles displayed on such prestigious web pages are also passed through the inspection department review the content to suit all levels, as well as not breaking the law. As a result, the web pages are a safe and abundant source of data, which is the premise for Vietnamese-Chinese bilingual data.

In addition, we also introduced a method of data collection based on the library name (JSOUP). We know of this library under the link in the reference [3] to get text inside the HTML[4] tag name. This open-source library is a Java library, created to work with HTML. This library provides a very convenient API[5] for extracting and manipulating data, using the best of DOM[6], CSS[7] and jQuery methods.

The uses of this library include: extracting, analyzing HTML from paths, finding and extracting data, manipulating elements, and preventing XSS[8] attacks.

First, we performed the method of finding the largest paging number to determine the total number of paging that can be retrieved and to save the current pagination count on the total pagination. From there, we proceed to configure to get the text in the title tag and content tag. After several improvements, we made the data acquisition rate considerably (In detail, in section 4 we present the experimental).

We have conducted data collection on some newspaper websites, with the data we collected in Section 4 (Experimental). We collect and save as text files, for easy processing and archiving. In addition, we developed the application to facilitate the collection of additional data from other websites and to retrieve new data from the collected websites.

The rest of this paper is structured as follows: Section 2 presents some related works, section 3 details the method for obtaining data, section 4 shows and discusses the results of our experiments. Finally, section 5 is the direction of development and conclusion.

2 Related work

In this section, we focus on survey methods of collecting data from the web. For more such automated data collection and collection methods so there are legitimate and know more open-source libraries that support get data from the web.

In the reference [1] have a research section on automatic data collection techniques that are legal in countries around the world. As a result, there have not been any legal proceedings related to automatic data extraction. However, in Germany, there was a decision on online data ownership and the prevention of automatic data collection.

[4] HTML: Hypertext Markup Language
[5] API: Application Program Interface
[6] DOM: Document Object Model
[7] CSS: Cascading Style Sheets
[8] XSS: Cross-Site Scripting

But most of the law in the world, just do not use tricks to collect accounts, passwords, personal information, or data files are prohibited download and do not damage the data owner. Before paying the fee, it is legal.

With the research in reference [1], we extracted the text automatically from the web pages, which is completely legal, because the network news is for everyone to read and be in public. In addition, we do not use any other tricks to obtain personal information from users or to damage the owner.

In the reference [2] research on extraction methods and text mining techniques, they have a section talking about open-source libraries for extracting data and giving out some open-source code. That includes JSOUP that we are using. Giulio Barcaroli [2] also found some difficulties in using this open-source website that is not fully accessible and not entirely based on standard HTML text.

From research [2] we know the difficulties we face and should depend on the technology used for that page, we have different ways of obtaining data.

3 Method for obtaining data

To facilitate the retrieval of data from various network sources, to have large experiment data sources when needed, we have created an application to retrieve data based on open-source library JSOUP, this open-source library support allows us to retrieve content inside HTML tags when we access that site. For example, with a newspaper web page we only need to get the title and content of that page, not the other sections, when using this open-source library, we put the input is the tag name contains the content to get in HTML with the path of that page, this open-source library will return us the content inside that tag.

We select bulletin pages with bilingual data as input, to get large data and accurate experiment environment the first thing is to find reputable news sites, high reliability. Then proceed to get all the articles (not including pictures and movies).

Because of the excessive number of articles on a page, to reduce the search space for articles with the same content, from the beginning, the data must be classified. For example, vietnamplus.vn has data that is in Vietnamese and in Chinese, we collect data classified by categories (politics, economics, culture, etc.).

With bilingual data Vietnamese-Chinese, these sites are mainly for users Vietnamese, so Vietnamese data more than Chinese data and Vietnamese data is often written more new information than Chinese data. So, the original sample data from the web pages were not balanced and “noise”, we must use many refinement methods were used to find pairs of sentences, words bilingual.

Each bilingual web page has different characteristics, even the same page, but the structure between Vietnamese and Chinese is different, or the same language on one page but the categories (economic, social, etc.) also different. For example, on the same page but the “economic” category for traditional paging and data storage (each category of information displays the last page number), but also the same page the "social" category is not paging and use method ajax data transmission (the user will read the latest news at the top, if the user wants to see the older news must pull web page down, the server will request the older news to return to the user).

Figure 1 illustrates the method of retrieving data. We chose the Java programming language to implement methods for retrieving data from the network. Also, using the JSOUP library to parse, separates the structure of an HTML page into tags created by the page development team, which then retrieves the content within the tag name that we want.

After the page selection steps to retrieve data and select content within the pages. We implement the following steps:

- Step 1: Determine the title and content of each page by tag in HTML.

We manually find the tag name of title and content in the HTML, then rely on Jsoup to retrieve the contents of the tag.

- Step 2- Find the largest pagination number in the category of the page to retrieve.

This step is very important when we have the largest paging we will take from the first paging to the last paging to ensure that the articles are retrieved. There is much data of some categories, some categories up to 200,000 articles in a category, so we do not always have time to get the full range from the first paging to the last paging (time to get data of a page up to several days), as well as prevent risks while retrieving data for disconnected phenomena, so we will save the current page on the total number of pages to we know how many percent data we have collected.

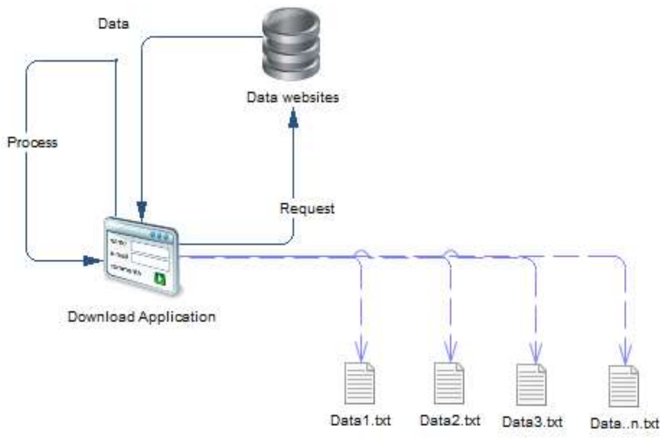


Figure 1. The structure of the download application

We implemented the way to get the largest paging by the pseudo code followings:

Algorithm:getMaxPagingNumber(String pageName)

*/ pageName is the path of the category we want to retrieve.

*/ result and factor are random numbers to start.

*/checkValidLink(result) is a function check with the paging path that exists or not.

```
Integer result = 1000, factor = 1000;
Boolean previous = true, reachInvalidLink = false;
while(true)
        if(checkValidLink(result))
                if(reachInvalidLink = true)
```

```
            factor = factor /2;
            result = result + factor;
        else
            result = result + factor;
            previous = true;
    else
        factor = factor/2;
        result = result - factor;
        previous = true;  reachInvalidLink = true;
    if(factor <= 5)
        break;
    if(factor >=5000 || factor <=0)
        return -1;
end while
if(pre = true){
*/ If you run 10000 times without results then stop
    for( i = result; i<10000; i++){
        if(checkValidLink(i))
            return i;
    end for
else
    for(i = result; i>=1; i--)
        if(checkValidLink(i))
            return i;
    end for
return -1;
```

Basically, we initially did not know what the maximum number of paging was. We give a random paging number to test, if that paging exists, continue to check for larger paging, and if that does not exist, we will check for smaller paging numbers until seeing that paging exists does not exist and no larger paging exists.

- Step 3- Get the title and content of the page and save the file. However, in the process of retrieving data may have many problems, the web page may exist without content or request timeout.

There are some web pages that exist but only the image with the name of the image, so the amount of data on the language in that category is very small. With this automated data collection, there will be pagings that do not exist during the process we start collecting from the first paging to the largest paging. So we will ignore all the risks, problems mentioned above to continue until the complete data of the web page.

4 Experiment

4.1 Toolkit in experiment.

We created our own application for the experiment based on open-source JSOUP on the Java programming language. With this open-source support us get the content inside the HTML tag, so we can configure each web page individually for content because each site has different types of configurations, in terms of our methods, we presented in section *3. Method for obtaining data.*

4.2 Characteristics of the web pages

Each site has different characteristics, but in terms of the amount of data collected, most web pages have more Vietnamese articles than Chinese articles, because the pages are of people Vietnam established. The specific speed of collection we have listed in section 4.3 Experiment results. So, we list the characteristics of the sites in Table 1 below.

Website address	Characteristics
http://www.vietnamplus.vn	-How to get paging number: The biggest paging is at the bottom of the web page of that category. Get the paging by JSOUP. Just run from the first paging to the last paging to retrieve the data. -During the scraping process, there are some pages that exist, are accessible to, but the system is under maintenance. -Many articles and data. -The speed when data extraction fast, but not stable, slow gradually.
http://baobinhduong.vn	- How to get paging number: The maximum number of paging is not available when the test number is bigger than the maximum paging, the site automatically returns the page with the biggest number paging. - Average data extraction speed and stable.
http://www.nhandan.com.vn	- How to get paging number: By algorithm in section 3. - Average data extraction speed and stable with Chinese but Vietnamese is very lower and lower in the end.

Table 1. Characteristics of the web pages that have collected the data.

4.3 Experiment result

We conduct the experiment on Ubuntu operating system version 16.04 LTS 64 bits with some configuration information such as 16GB RAM, Intel Core i5-7400 CPU, GeForce GT 730/PCIe/SSE2 graphics and stable network connection 4.1 MB/s download speed and 4.7 MB/s upload speed.

All of our experiment web pages are dissimilarity to categories. In the tables in this section, the numbers **0** appear in the table are no corresponding content to collect data.

From the statistical results in Table 2 and Table 3 below, the bilingual data of the website http://www.vietnamplus.vn is quite numerous in both Chinese and Vietnamese. The number of Vietnamese files is more than six times that of Chinese, and the total size of files is seven times as many. While Vietnamese has 3 categories which are not available in Chinese,

Chinese, vice versa, contains 1 of those which do not appear in Vietnamese.

Vietnamese		Chinese	
The number of files obtained	Total size (MB)	The number of files obtained	Total size (MB)
49,031	231	11,631	43.5
12,808	68.6	15,287	62.1
24,472	108	766	2.94
15,868	77.4	5,727	21.4
22,721	94	928	3.56
37,717	184	2,545	9,85
114,432	456	960	4.01
11,707	55,5	0	0
8,896	37.6	0	0
3,686	5.32	0	0
0	0	1,147	4.35

Table 2. Result of bilingual data collection was categorized according to the content of the website http://www.vietnamplus.vn

http://www.vietnamplus.vn	Number of categories	The number of files obtained	Total size (MB)	Download speed (files/minute)
Chinese	9	46,863	181	434
Vietnamese	10	301,422	1351	273

Table 3. Results of bilingual data collection page http://www.vietnamplus.vn

http://baobinhduong.vn	Vietnamese		Chinese	
	The number of files obtained	Total size (MB)	The number of files obtained	Total size (MB)
Economy	13,874	78.9	6,317	24.4
Politic	18,516	77	2,711	5.10
Health	60	0.3	1,021	1.48
Cultural	5,439	26,8	2,114	8.10
Technology	60	0.3	496	1.92
Society	15,701	80.9	5,567	20.9
International	8,257	37.1	1,553	5.96
Environment	1,979	9,78	324	1.25
Policy	**0**	**0**	335	1.26
Real estate	**0**	**0**	97	0.39
Sea Island	**0**	**0**	110	0.4
Travel	**0**	**0**	968	3.71
Vietnam and the World	**0**	**0**	2,751	10.8
You read	2,139	12.9	**0**	**0**
Enterprise	933	5.64	**0**	**0**
Family	2,755	13.7	**0**	**0**
Profile Documentation	1,057	10.2	**0**	**0**
Labor	2,394	12.5	**0**	**0**
Car and Moto	60	0.3	**0**	**0**
Analysis	2,089	11.1	**0**	**0**
Law	4,837	24.1	**0**	**0**
Defence Security	2,302	12.1	**0**	**0**
Sport	8,192	40.6	**0**	**0**

Table 4. Result of bilingual data collection was categorized according to the content of the website http://baobinhduong.vn

http://baobinhduong.vn	Number of categories	The number of files obtained	Total size (MB)	Download speed (files/minute)
Chinese	13	24,344	39.3	286
Vietnamese	18	90,644	492	192

Table 5. Results of bilingual data collection page http://baobinhduong.vn

The statistical results in Table 4 and Table 5 show that the bilingual data of http://baobinhduong.vn is much different. The number of files collected in Vietnamese is more than 4 times, but the total file size of Vietnamese is more than 12 times. Therefore, we anticipate that the content within the same Chinese files on one content will be far less than the Vietnamese one. While Vietnamese has 10 categories which are not available in Chinese, Chinese, vice versa, contains 5 of those which do not appear in Vietnamese.

http://www.nhandan.com.vn	Vietnamese		Chinese	
	The number of files obtained	Total size (MB)	The number of files obtained	Total size (MB)
Economy	27,324	163	6,287	22.2
Politic	57,256	345	6,701	26.8
Society	35,507	183	6,861	24.5
Cultural	18,663	111	1,762	6.14
World	34,258	154	4,143	15.2
Sport	8,567	39.4	1,532	5.02
You read	5,136	25.3	**0**	**0**
Technology	4,761	23.2	**0**	**0**
Education	8,126	47.5	**0**	**0**
Science	6,652	39.3	**0**	**0**
Law	17,438	114	**0**	**0**
Health	7,436	39.9	**0**	**0**
Asian	**0**	**0**	252	1.01
Document	**0**	**0**	244	1.33
Friendship bridge	**0**	**0**	5,940	22.2
Leader	**0**	**0**	2,320	9.3
Internal	**0**	**0**	287	1.01
Vietnam country	**0**	**0**	77	0.3
Travel	**0**	**0**	1,690	5.87
Vietnam window	**0**	**0**	202	0.8

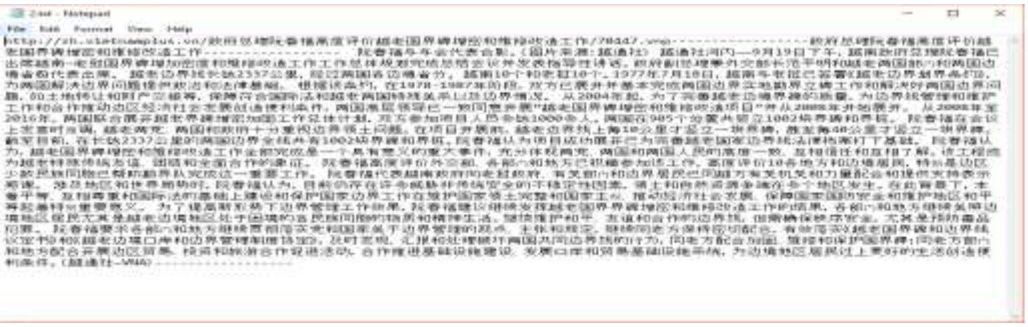
Table 6. Result of bilingual data collection was categorized according to the content of the website http://www.nhandan.com.vn

http://www.nhandan.com.vn	Number of categories	The number of files obtained	Total size (MB)	Download speed (files/minute)
Chinese	14	38,298	142	103
Vietnamese	12	231,133	1280	110

Table 7. Results of bilingual data collection page http://www.nhandan.com.vn

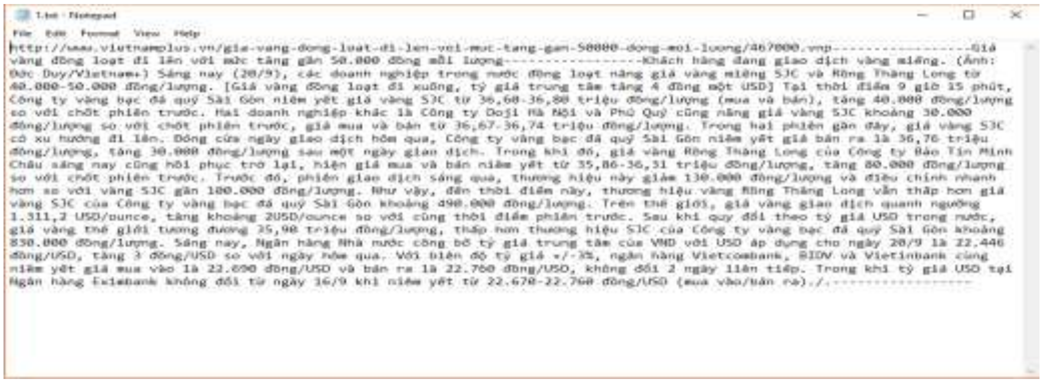
The statistical results in Table 6 and Table 7 show that the bilingual data of the site http://www.nhandan.com.vn is much different in terms of the number of files received, the total size of the files. The categories with the same overlapping content are 6 out of 20. The Chinese version of this article is pretty much in bilingual but not as much in Vietnamese, but this is also a potential site of extraction, analysis later for the bilingual problem.

As shown in Figure 2 and Figure 3, we save the data by the text file and save the link, title, and content of the article, each content is separated by the string "------------------------ ", to facilitate the analysis process, extracted bilingual data.



http://zh.vietnamplus.vn/[illegible]

Figure 2. Chinese data after collection.



http://www.vietnamplus.vn/gia-vang-dong-loat-di-len-voi-muc-tang-gan-50000-dong-moi-luong/467000.vnp-----------------Giá vàng đồng loạt đi lên với mức tăng gần 50.000 đồng mỗi lượng-----------------Khách hàng đang giao dịch vàng miếng. (Ảnh: Đức Duy/Vietnam+) Sáng nay (20/9), các doanh nghiệp trong nước đồng loạt nâng giá vàng miếng SJC và Rồng Thăng Long từ 40.000-50.000 đồng/lượng. [Giá vàng đồng loạt đi xuống, tỷ giá trung tâm tăng 4 đồng một USD] Tại thời điểm 9 giờ 15 phút, Công ty vàng bạc đá quý Sài Gòn niêm yết giá vàng SJC từ 36,60-36,80 triệu đồng/lượng (mua và bán), tăng 40.000 đồng/lượng so với chốt phiên trước. Hai doanh nghiệp khác là Công ty Doji Hà Nội và Phú Quý cũng nâng giá vàng SJC khoảng 30.000 đồng/lượng so với chốt phiên trước, giá mua và bán từ 36,67-36,74 triệu đồng/lượng. Trong hai phiên gần đây, giá vàng SJC có xu hướng đi lên. Đóng cửa ngày giao dịch hôm qua, Công ty vàng bạc đá quý Sài Gòn niêm yết giá bán ra là 36,76 triệu đồng/lượng, tăng 30.000 đồng/lượng sau một ngày giao dịch. Trong khi đó, giá vàng Rồng Thăng Long của Công ty Bảo Tín Minh Châu sáng nay cũng hồi phục trở lại, hiện giá mua và bán niêm yết từ 35,86-36,31 triệu đồng/lượng, tăng 80.000 đồng/lượng so với chốt phiên trước. Trước đó, phiên giao dịch sáng qua, thương hiệu này giảm 130.000 đồng/lượng và điều chỉnh nhanh hơn so với vàng SJC gần 100.000 đồng/lượng. Như vậy, đến thời điểm này, thương hiệu vàng Rồng Thăng Long vẫn thấp hơn giá vàng SJC của Công ty vàng bạc đá quý Sài Gòn khoảng 490.000 đồng/lượng. Trên thế giới, giá vàng giao dịch quanh ngưỡng 1.311,2 USD/ounce, tăng khoảng 2USD/ounce so với cùng thời điểm phiên trước. Sau khi quy đổi theo tỷ giá USD trong nước, giá vàng thế giới tương đương 35,90 triệu đồng/lượng, thấp hơn thương hiệu SJC của Công ty vàng bạc đá quý Sài Gòn khoảng 830.000 đồng/lượng. Sáng nay, Ngân hàng Nhà nước công bố tỷ giá trung tâm của VND với USD áp dụng cho ngày 20/9 là 22.446 đồng/USD, tăng 3 đồng/USD so với ngày hôm qua. Với biên độ tỷ giá +/-3%, ngân hàng Vietcombank, BIDV và Vietinbank cùng niêm yết giá mua vào là 22.690 đồng/USD và bán ra là 22.760 đồng/USD, không đổi 2 ngày liên tiếp. Trong khi tỷ giá USD tại Ngân hàng Eximbank không đổi từ ngày 16/9 khi niêm yết từ 22.670-22.760 đồng/USD (mua vào/bán ra)./.-----------------

Figure 3. Vietnamese data after collection.

5 Conclusion

We have now successfully experimented 7 bilingual sites. In order to have a good input source for bilingual data, we will be experiments on more web pages with bilingual data. In addition, we will provide custom instructions to the community, if anyone wants to retrieve data from any web page they can manually customize and retrieve data based application backgrounds we have created available.

In this paper, we were given the method how to get data based on an open source library called JSOUP. This automated data retrieval system is suitable for translation teams that do not have language experts or want to reduce time and costs, but the input data will not be as perfect as a Vietnamese file and a Chinese file translation, imperfect data problem is unavoidable.

With the experimental results, we plan to select a few pages with Chinese and Vietnamese content with not too much difference in the number of files and the size of the content to serve as input to the analysis, filter out the files, articles with similar content, from which we can collect bilingual data.

References

[1] *Automatic data collection on the Internet (web scraping)* VERSION 18 May 2015

[2] Giulio Barcaroli, Alessandra Nurra, Marco Scarnò, Donato Summa Istituto Nazionale di Statistica Cineca, *Use of web scraping and text mining techniques in the Istat survey on “Information and Communication Technology in enterprises”*

[3] *https://jsoup.org/*

[4] Charmaine Bonifacio, Thomas E. Barchyn, Chris H. Hugenholtz, Stefan W. Kienzle, *CCDST: A free Canadian climate data scraping tool*

[5] Salim Khalil, Mohamed Fakir, *RCrawler: An R package for parallel web crawling and scraping*

[6] Lasse Johansson, Victor Epitropou, Kostas Karatzas, Ari Karppinen, Leo Wanner, Stefanos Vrochidis, Anastasios Bassoukos, Jaakko Kukkonen, Ioannis Kompatsiaris, *Fusion of meteorological and air quality data extracted from the web for personalized environmental information services.*