Automatic Searching for English-Vietnamese Documents on the Internet

Quoc Hung Ngo[1], *Dinh Dien*[2], *Werner Winiwarter*[3]
(1) Faculty of Computer Science, University of Information Technology, HoChiMinh City, Vietnam
(2) Faculty of Information Technology, University of Natural Sciences, HoChiMinh City, Vietnam
(3) University of Vienna, Research Group Data Analytics and Computing, Vienna, Austria
hungnq@uit.edu.vn, ddien@fit.hcmus.edu.vn,
werner.winiwarter@univie.ac.at

ABSTRACT

Bilingual corpora together with machine learning technology can be used to solve problems in natural language processing. In addition, bilingual corpora are useful for mapping linguistic tags of less popular languages, such as Vietnamese, and for studying comparative linguistics. However, Vietnamese corpora still have some shortcomings, especially English–Vietnamese bilingual corpora. This paper focuses on a searching method for bilingual Internet materials to support establishing an English–Vietnamese bilingual corpus. Based on the benefit of natural language processing toolkits, the system concentrates on using them as a solution for the problem of searching any Internet English–Vietnamese bilingual document without the need for any rules. We propose a method for extracting the main content of webpages without the need for frame of website or source of website before processing. Several other natural language processing tools included in our system are English-Vietnamese machine translation, extracting Vietnamese keywords, search engines, and comparing similar documents. Our experiments show several valuable auto-searching results for the US Embassy and Australian Embassy websites.

ABSTRACT (L_2)

Sự kết hợp giữa ngữ liệu song ngữ và máy học có thể giúp giải quyết nhiều vấn đề trong xử lý ngôn ngữ tự nhiên. Hơn nữa, các ngữ liệu song ngữ còn giúp ích rất nhiều trong việc ánh xạ nhãn ngôn ngữ cho các ngôn ngữ ít phổ biến như tiếng Việt và các nghiên cứu trong ngôn ngữ học so sánh. Tuy nhiên, ngữ liệu tiếng Việt vẫn còn có ít và hạn chế, đặc biệt là ngữ liệu song ngữ Anh-Việt. Bài báo này tập trung vào việc đưa ra một phương pháp tìm kiếm các tài liệu song ngữ từ nguồn dữ liệu Internet nhằm hỗ trợ cho việc xây dựng bộ ngữ liệu song ngữ Anh-Việt. Dựa trên những công cụ xử lý ngôn ngữ tự nhiên hiện có, hệ thống tập trung vào sử dụng chúng như là một giải pháp để giải quyết vấn đề tìm kiếm tài liệu song ngữ bất kỳ từ Internet mà không cần các quy tắc định trước. Chúng tôi cũng đề xuất một phương pháp để rút trích nội dung chính trang web mà không phải phụ thuộc vào thiết kế cũng như nguồn gốc của trang web đó. Một số công cụ xử lý ngôn ngữ tự nhiên khác sử dụng trong hệ thống của chúng tôi gồm có dịch máy tự động Anh-Việt, rút trích từ khóa tiếng Việt, tìm kiếm tài liệu từ Internet, và so sánh độ tương đồng văn bản. Bài báo cũng đưa ra một số thử nghiệm với các kết quả có giá trị trong quá trình tìm kiếm tự động các tài liệu song ngữ từ nguồn là trang web của Đại sứ quán Hoa Kỳ và Đại sứ quán Úc.

Keywords: Building Corpus, English-Vietnamese Bilingual Corpus, Search Engines
Keywords (L_2): Xây dựng Ngữ liệu, Ngữ liệu Song ngữ Anh-Việt, Công cụ Tìm kiếm

1 Introduction

Nowadays, corpus-based NLP research has been developing rapidly. There are many corpus-based studies and tools in machine translation (Koehn, 2005), information retrieval (Chen, 2000; Hawking, 1996; Oard, 1998), bitext alignment (Burkett, 2010), etc. These tools are built flexibly to work on many languages with an input corpus. For example, YAMCHA toolkit[1] was built for classifying tasks, such as POS tagging, Named Entity Recognition, and Text Chunking. It works effectively on English partly because of the quality of English corpora. However, its result on Vietnamese is not as good as on English because of the low quality and quantity of Vietnamese corpora. Moreover, bilingual corpora are also used to map linguistic tags from English to other languages (Dien, 2003). Hence, building monolingual and bilingual corpora is still valuable for many languages.

With the vast resources from the Internet, many researchers have studied to mine them to build bilingual corpora, such as Yang, C.C. and co-authors (Yang, 2003) and Zhang, Y. and co-authors (Zhang, 2006) for the English-Chinese pair; and Van, D.B. and Quoc, H.B. (Van, 2007) and Vu, P.D.M. (Vu, 2007) for the English-Vietnamese pair. However, their bilingual corpus building systems work on the supposition that these documents and their translations come from the same origin. It means that their URL addresses have the same domain or at least related domains. For the English-Vietnamese pair, authors used a bilingual dictionary to look up the meaning of words in the English documents, then search translation documents from a specific domain which cover these word meanings.

This project points out a system to search English-Vietnamese bilingual documents on the Internet by combining several NLP modules: extracting keywords, machine translation, search engine, and comparing similar documents. The system is based on a framework-free web content extraction module and search engines, therefore, it leads to our system being independent from the domains in which it searches the candidates of translation documents.

2 Related work

2.1 WPDE system

The WPDE system of Zhang and co-authors has three main stages: choosing candidate websites and extracting web contents; extracting parallel bilingual document pairs; and analyzing translation pairs (Zhang, 2006). Features which are used for choosing bilingual document pairs are domains in URLs, addresses and filename. The system determines such words, phrases such as "e", "en", "eng", "english" for English and "c", "ch", "chi", "cn" or "zh" for Chinese. Moreover, the system also uses the similarity in the html structure to detect translation candidate pairs. Analyzing the translation pairs is based on mechanical features, such as file size between two files, structures of html pages, etc. Finally, the WPDE's model uses the k-nearest neighbor approach to classify candidate pairs.

2.2 PTMiner System

PTMiner system of Chen, J., Nie, J.J. searches and identifies English–Chinese bilingual sites automatically (Chen, 2000). The system works in a similar way as Resnik's system (Resnik,

[1] http://chasen.org/~taku/software/yamcha/

1998). However, the authors used several features for filtering and detecting bilingual websites after downloading and used an approach based on web content and content alignment:

+ Filtering based on length of web pages,
+ Filtering based on structures,
+ Filtering based on content alignment.

2.3 English-Vietnamese Alignment System

The system of Van D.B. and Quoc H.B. (Van, 2007) and the system of Vu P.D.M. (Vu, 2007) download web pages from specific addresses or domains (such as www.voanews.com). In the same way as other systems, they remove HTML tags and get the main content as the raw document of the process. Then, they use several heuristics to detect parallel bilingual pairs. These candidates are analysed by a two-step filter:

+ Sentence length filtering
+ Lexicon-based sentence alignment

To reduce the dictionary size and increase precision, the authors used the Porter algorithm to stem English words before looking them up in the dictionary (Van, 2007). The authors applied their approach on the VOANews website, and the precision was 90% with about 8,500 bilingual sentence pairs. However, the limitation of the approach is using heuristics to create candidate pairs.

3 Research framework

The system uses an English–Vietnamese translation system, a keyword extraction tool based on Vietnamese words, and a document similarity comparison tool to find translated documents of English pages. Translated pages are found on the Internet by using the Google Search engine and NLP toolkits. Our searching system includes three main phases: (1) download webpages and get web content; (2) translate and search candidates; and (3) compare similarities between translated document and candidates (see Figure 1).

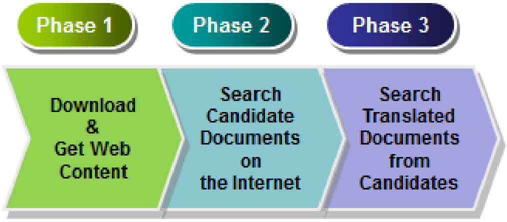

Phase 1
Phase 2
Phase 3
Download & Get Web Content
Search Candidate Documents on the Internet
Search Translated Documents from Candidates

FIGURE 1 – Architure of the system for searching English-Vietnamese documents automatically

In general, Phase 1 includes downloading webpages, parsing the HTML files into the HTML trees, and then choosing the main content from the tree. Phase 2 includes translating main content of webpages into Vietnamese text by using Google Translate[2], extracting keywords from Vietnamese text, and searching Vietnamese documents on the Internet by extracted keywords

[2] http://translate.google.com/

and saving them as candidate documents. Finally, Phase 3 is simply comparing similarity between the translated Vietnamese document and candidate documents.

3.1 Get web content

Because results of the searching system can be from any resource domain, the first task of the system is detecting the main content of web pages of any domain which can be unknown in advance. These web pages are in English or Vietnamese. Unlike the framework-based approaches (Vu, 2007), our approach for this task is parsing web pages into HTML trees, identifying the content node in these trees, and then analysing their contents (see Figure 2 and Figure 3).

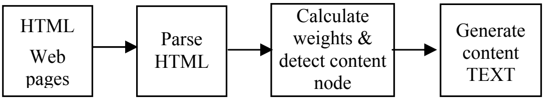
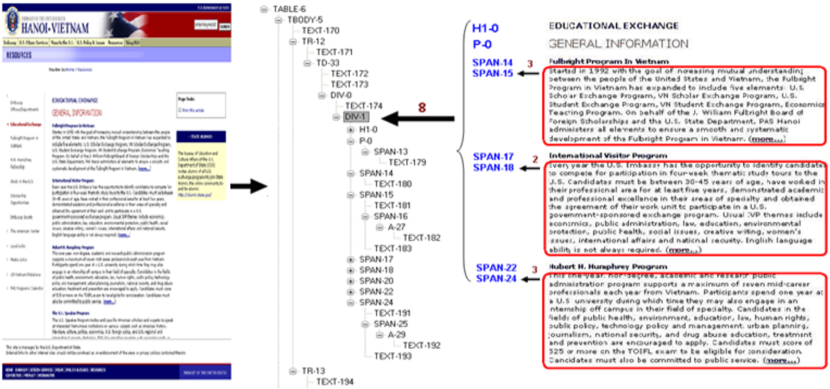


FIGURE 2 – Model of extracting web content

Parsing the web page into an HTML tree is based on the Majestic12 project[3] . Next, the process of extracting the main content includes two steps: analysing and giving marks to HTML nodes and then detecting the content node in the HTML tree. Giving marks to HTML nodes is based on counting content sentences for these nodes. Rules for calculating weights for nodes in HTML trees are (see an illustration in Figure 3):



EDUCATIONAL EXCHANGE

GENERAL INFORMATION

Fulbright Program In Vietnam

Started in 1992 with the goal of increasing mutual understanding between the people of the United States and Vietnam, the Fulbright Program in Vietnam has expanded to include five elements: U.S. Scholar Exchange Program, VN Scholar Exchange Program, U.S. Student Exchange Program, VN Student Exchange Program, Economics Teaching Program. On behalf of the J. William Fulbright Board of Foreign Scholarships and the U.S. State Department, PAS Hanoi administers all elements to ensure a smooth and systematic development of the Fulbright Program in Vietnam. (more...)

International Visitor Program

Every year the U.S. Embassy has the opportunity to identify candidates to compete for participation in four-week thematic study tours to the U.S. Candidates must be between 30-45 years of age, have worked in their professional area for at least five years, demonstrated academic and professional excellence in their areas of specialty and obtained the agreement of their work unit to participate in a U.S. government-sponsored exchange program. Usual IVP themes include economics, public administration, law, education, environmental protection, public health, social issues, creative writing, women's issues, international affairs and national security. English language ability is not always required. (more...)

Hubert H. Humphrey Program

This one-year, non-degree, academic and research public administration program supports a maximum of seven mid-career professionals each year from Vietnam. Participants spend one year at a U.S. university during which time they may also engage in an internship off campus in their field of specialty. Candidates in the fields of public health, environment, education, law, human rights, public policy, technology policy and management, urban planning, journalism, national security, and drug abuse education, treatment and prevention are encouraged to apply. Candidates must score of 525 or more on the TOEFL exam to be eligible for consideration. Candidates must also be committed to public service. (more...)

FIGURE 3 – Illustration of extracting web content for an US Embassy web page.

+ Only weighing nodes with TEXT tags because these nodes contain real content.
+ Weights of TEXT nodes are based on the number of their sentences. The more sentences nodes have, the higher weights are.
+ Nodes contain at least one paragraph.
+ Weights of parent nodes are calculated by summing all weights of children.

[3] http://www.majestic12.co.uk/projects/html_parser.php

Finally, the content node is the deepest node with the highest weight in the HTML tree. This approach can detect the content node and extract the main content of webpages without the need for frame of website as well as source of website before processing.

For example, node DIV-1 in Figure 3 contains node H1-0, P-0, SPAN-14, SPAN-15, SPAN-17, SPAN-18, SPAN-20, SPAN-22, and SPAN-24. Weight of node SPAN-15, SPAN-18, and SPAN-24 are 3, 2, and 3, respectively, while weight of other nodes is zero because they contain phrases instead of complete sentences. Hence, weight of node DIV-1 is 8 (= 3+2+3).

3.2 Get keywords for searching

Unlike English, Vietnamese words can be a group of several tokens, therefore, extracting Vietnamese keywords has to consider to extract words instead of tokens. Firstly, word segmentation is implemented by ensuring that extracted keywords have meaning and are real words. Steps for extracting Vietnamese keywords include:

+ Segment Vietnamese words,

+ Remove Stopwords,

+ Estimate term frequencies.

Calculating and extracting keywords has two sub-steps (Matsuo, 2004):

+ Calculate the co-occurrence frequency of word *w* and word *g* by *freq(w,g)*

+ Calculate co-occurrence with the $\chi 2$ estimate.

Keywords are words which have highest $\chi 2$ weights in the document. In our experiment, the system only extracts 3 first keywords for next searching step (based on our experiment shown in Section 4.2).

3.3 Using search engines for searching candidates

Keywords are provided for the searching system by generating an address which includes these keywords. Searching result is search engines' response file which is acquired by this address. For example, keywords "***building parallel corpus***" generate URL addresses for searching systems:

Google: http://www.google.com.vn/search?hl=vi&q=building+parallel+corpus

Yahoo: http://search.yahoo.com/search?p=building+parallel+corpus

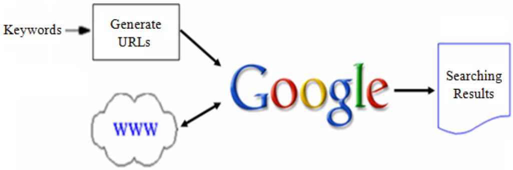


FIGURE 4 – Searching process by Google system

By acquiring web page of these addresses, the system gains searching results from Google or Yahoo (as shown in Figure 4). Google and Yahoo also support searching results which come from any domain, a specific domain or except a specific domain. This option allows our auto-searching system to be able to direct its searching results. It can be implemented by adding following parameters:

- Search in specific domain: ***site:<domain>***
- Search in except domain: ***-site:<domain>***

For example, the URL address for searching "*building parallel corpus*" from site www.aclweb.org is "site:aclweb.org building parallel corpus". It means that every result of the search engine will come from aclweb.org.

From Google's results, the system continues to download candidate documents from the result addresses and extracts main content to store them as candidates for evaluating translation pairs at phase 3, comparing similarity between the automatic translation document and candidates.

3.4 Document comparison for Vietnamese

In general, the comparison module includes two steps: mechanical filter and content-based similarity comparison. The mechanical filter uses several features to remove less suitable candidates, such as rate of file size between two documents, the differences in number of paragraphs or sentences. The content-based similarity comparison step calculates documents as vectors. Words which are chosen for representation in the vector are the top 30% frequent words in the document. Comparing two documents becomes calculating the distance between two vectors by measuring the Euclidean distance between two document vectors (Guo, 2008).

4 Experiments and results

4.1 Get web content

In general, getting web content works effectively on embassy websites and result webpages of the searching process. Table 1 shows the result of getting web content from the US Embassy website and searching results from the Internet (from unanticipated domains).

For news expresses, the number of articles is very huge and there is additional information, such as title, abstract, category, publish date, authors, etc. This information is very useful for other natural language processing tasks which are based on this corpus. Hence, the system builds several specific definitions to identify content nodes for these websites. These definitions help the system to be able to extract effectively and get more information fields.

	USEmbassy	**Other Domains**[4]
Number of web pages	2,054	3,973
Number of main content files	1,870	3,035
Correct result	1,853	2,885
	99%	95%

TABLE 1 – Result of extracting main content

4.2 Keywords for search engines

Results of search engines depend on the provided keywords. To know how many keywords are

[4] Statistics based on 3,973 return records of the search engine from out of the US. Embassy website.

good for the bilingual document searching system, this project implements an experiment on the data from the Australian embassy website with 86 parallel document pairs. Tests acquire 30 returns from the Google search engine. The test has a different number of keywords to calculate the precision of returns of the search engine for our system. The result is shown in Table 2.

Number of keywords	2	3	4	5	6
Results of Google search	1,652	1,419	1,393	1,174	1,118
Result in parallel pairs	37	45	35	22	13

TABLE 2 – Results of Google search with several different numbers of keywords

Our experiments show that the accuracy of the searching module is highest when looking for 3 keywords for each document (see Table 2). In this test, correct results of search engines are addresses in the parallel pairs. In fact, the Google's response for 3 keywords has 1,419 results with 45 correct results, while it has 1,652 results with 37 correct results for 2 keywords. Similarly, the Google search just acquires 35, 22 and 13 correct results when searching with 4, 5, and 6 keywords, respectively.

4.3 Searching results on embassy websites

In our first experiment, the system carries out several tests on the US Embassy and Australian Embassy websites with totally 2,500 web pages in which there are 440 parallel English–Vietnamese document pairs (see Table 3).

Website	URL	Content pages	Data size
USEmbassy: English site	896	832	2.4 MB
USEmbassy: Vietnamese site	1,150	1,076	12.8 MB
AUEmbassy: English site	207	178	2.4 MB
AUEmbassy: Vietnamese site	158	128	12.8 MB

TABLE 3 – Characteristics of data from embassy websites

Table 4 shows the result of bilingual website of the US Embassy and Australian Embassy with the similarity threshold of 0.7 .

Website	Parallel pairs	Results	Precision (%)	Recall (%)
USEmbassy	354	197	92%	51%
AUEmbassy	86	45	93%	49%

TABLE 4 – Result of the bilingual document searching system on the US Embassy and Australian embassy pages after Phase 1

Table 5 shows the results of whole system when applying it to the US Embassy and Australian Embassy bilingual websites.

Website	Parallel pairs	Results	Precision (%)	Recall (%)
USEmbassy	354	350	90%	89%
AUEmbassy	86	84	92%	90%

TABLE 5 – Result of the searching bilingual document system on the US Embassy and Australian Embassy pages after Phase 2

In our experiments, recall parameter is only evaluated on the webpages of the US Embassy and Australian Embassy instead of on the global Internet because the resources from the Internet are enormous and, maybe, there are some correct pages from other domains which our system does not discover.

4.4 Searching result on other websites

Result of searching translation documents depends on the content of web pages. Translation documents come from any domain in the Internet. However, the original documents which are translated and then posted on other domains as well as re-published as translation versions on other locations mainly come from the Press Releases division of the Embassy website.

Another experiment is focused on news announcements of the US Embassy. It has 354 English news articles in the Press Releases section of the US Embassy website while only some articles are translated into Vietnamese and posted on the Embassy website (see Table 6).

Website	English pages (Press Releases)	Vietnamese documents in searching results
USEmbassy	354	52

TABLE 6 – Result of searching parallel documents from embassy websites

In 52 found results, there are ten results for which there are no translations on the US Embassy website. These articles are news announcements which were published in English by the US Embassy before 2004 (without in Vietnamese) and which were translated and published on News expresses.

5 Conclusions

The project has pointed out a process to search English—Vietnamese bilingual documents on the Internet. The approach allows to look for translation documents from any resource instead of from a specific domain. We also demonstrated the process by building an application to search bilingual documents automatically, the result on the US Embassy and Australian Embassy websites are very promising. Moreover, the project also has produced several valuable tools, such as getting web content, extracting Vietnamese keywords, comparing similar documents. Particularly, for getting main content, we presented a method to calculate content nodes in the HTML tree and extract the main content from the HTML parser result. This tool can be used for NLP projects which are based on web contents.

6 References

Burkett, D., Petrov, P., Blitzer, J., and Klein, D. (2010). Learning Better Monolingual Models with Unannotated Bilingual Text. Proceedings of CoNLL 2010, pp. 46-54.

Chen J., and Nie, J. (2000). Parallel Web Text Mining for Cross-Language IR. Proceedings of RIAO-2000: Content-Based Multimedia Information Access. College de France, Paris, France, pp. 188-192.

Dien, D., and Kiem, H. (2003). POS-tagger for English-Vietnamese bilingual corpus, Proceedings of HLT-NAACL, Workshop on Building and Using Parallel Texts, pp. 88-95.

Guo, Q. (2008). The Similarity Computing of Documents Based on VSM, 32nd Annual IEEE International Computer Software and Applications Conference, pp. 585-586.

Koehn, Ph. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation, In Proceedings of the 10th Machine Translation Summit, Phuket, Thailand, pp. 79-86.

Hawking, D., Bailey, P., and Campbell, D. (1996). A Parallel Document Retrieval Server for the World Wide Web. Proceedings of the Australian Document Computing Symposium (ADCS), Melbourne, pp. 73-78.

Matsuo, Y., and Ishizuka, M. (2004). Keyword Extraction From A Single Document Using Word Co-Occurrence Statistical Information, International Journal on Artificial Intelligence Tools, Vol. 13, No. 1, pp. 157-169.

Oard, D. W., and Diekema, A. R. (1998). Cross-Language Information Retrieval. Annual Review of Information Science and Technology (ARIST), Volume 33, pp. 223-256.

Resnik, P. (1998). Parallel Stands: A preliminary investigation into mining the web for bilingual text. Proceedings of AMTA’98, pp. 72-82.

Van, D.B., and Quoc, H.B. (2007). Automatic Construction of English-Vietnamese Parallel Corpus through Web Mining, In Proceedings of 5th IEEE International Conference on Computer Science - Research, Innovation and Vision of the Future (RIVF’2007), Hanoi, Vietnam, pp. 261-266.

Vu, P.D.M. (2007). Building bilingual corpus by mining data from Internet, Master thesis in Computer Science, University of Natural Science, HoChiMinh City, Vietnam, 2007 (in Vietnamese).

Yang, C.C., and Li, K.W. (2003). Automatic Construction of English/Chinese. Parallel Corpora, Journal of the American Society for Information Science and Technology, pp. 730-742.

Zhang, Y., Wu, K., Gao, J., and Vines, P. (2006). Automatic Acquisition of Chinese–English Parallel Corpus from the Web. Advances in Information Retrieval, Volume 3936/2006, pp. 420-431.