**Matching document from very non-parallel corpus**

Minh Trinh1, Quang Le2, Phuoc Tran3

*NLP-KD Lab, Faculty of Information Technology,*

*Ton Duc Thang University, Ho Chi Minh city, Vietnam*

*1Email:trinhvanminh2009@gmail.com*

*2Email:qquangz006@gmail.com*

*3Email: tranthanhphuoc@tdt.edu.vn*

**Tóm tắt: Ngữ liệu đơn - song ngữ là vô cùng cần thiết, phục vụ cho các bài toán xử lý ngôn ngữ tự nhiên, đặc biệt là dành cho dịch máy. Hoa-Việt là cặp ngôn ngữ nghèo tài nguyên, dữ liệu song ngữ Hoa-Việt hiện còn hạn chế.Trong bài báo này, chúng tôi đưa ra các phương pháp phân tích để tìm ra các cặp file tương đồng từ dữ liệu không tương đồng mà chúng tôi đã hoàn thành bước thu thập tự động các cặp văn bản song ngữ Hoa-Việt từ các website song ngữ Hoa-Việt.**

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

.

1. Introduction

Kho ngữ liệu song ngữ là bắt buộc đối với các hướng tiếp cận dịch máy hiện đại, như dịch máy dựa trên thống kê hay dịch máy dựa trên nơ ron. Tất cả các phương pháp dịch máy đều được ước tính bằng cách huấn luyện tối đa và có khả năng lặp đi lặp lại dựa trên dữ liệu lớn các văn bản song ngữ tự nhiên. Từ đó cho thấy dữ liệu đầu vào để cho dịch máy thống kê là rất quan trọng và chất lượng của dịch máy cũng phụ thuộc vào dữ liệu. Do đó, xây dựng một tập hợp các câu, từ song ngữ là một nhiệm vụ hết sức quan trọng trong việc xây dựng bất kỳ hệ thống dịch máy nào.

Để có được dữ liệu cho dịch máy, các nhà nghiên cứu ngôn ngữ học thường sử dụng một trong hai phương pháp đó là thu thập thủ công hoặc rút trích văn bản tự động từ các website. Phương pháp thủ công có độ chính xác cao, nhưng phải đòi hỏi có một đội ngũ chuyên gia về ngôn ngữ học và am hiểu về song ngữ. Phương pháp thủ công thì tốn rất nhiều thời gian và chi phí.

Hiện nay với sự phát triển của các website song ngữ ngày càng phong phú cho nên phương pháp rút trích tự động cũng mang lại hiệu quả cao, tuy nhiên mức độ chính xác sẽ không cao bằng phương pháp thủ công, nhưng cũng có thể sử dụng trong nghiên cứu được. Với phương pháp tự động này sẽ tiết kiệm được thời gian và chi phí và số lượng từ ngữ thu được sẽ rất nhiều, nhưng phải xử lý và phân tích rất nhiều để có được kho dữ liệu song ngữ.

Trong bài báo này, mục đích của chúng tôi là hiện thực bước tìm kiếm các tập tin song ngữ sau khi đã hoàn thành bước thu thập dữ liệu tự động từ các website song ngữ Hoa-Việt. Từ kết quả của việc tìm ra các tập tin song ngữ có cùng nội dung là tiền đề cho bước rút trích các cặp câu song ngữ để tạo kho dữ liệu song ngữ Hoa-Việt.

Dữ liệu trước khi đưa vào xử lý đã được phân loại từ giai đoạn đầu, nên cũng giảm thiểu không gian xử lý. Do vậy, chúng tôi chỉ xử lý trên những tập dữ liệu có cùng nội dung như là chính trị, kinh tế, xã hội,… cùng trên một trang báo. Nhưng không thể có một bài báo tiếng Việt có một bản dịch sẵn ở bài báo tiếng Hoa, do các đội viết báo giữa các ngôn ngữ trên cùng một trang báo là hoàn toàn khác nhau (tất cả các trang báo có song ngữ Hoa-Việt mà chúng tôi biết đến). Do đó mục tiêu của chúng tôi đưa ra là tìm ra các tập tin có độ tương đồng cao nhất dựa trên các tiêu chí mà chúng tôi đưa ra.

Các tiêu chí mà chúng tôi đưa ra trong bài báo này là so sánh tiêu đề của các bài báo với nhau. Rút trích tập hợp thời gian bao gồm ngày, tháng, năm và tháng, năm (nếu có). Ngoài ra chúng tôi còn dựa trên tiêu chí so sánh các tập hợp các số phần trăm(nếu có). Chúng tôi so sánh trên một tiêu chí trước, sau đó nếu các tập tin nào có các tiêu chí như ngày, tháng, năm hoặc tiêu chí phần trăm có thì chúng tôi sẽ so sánh song song với nhau, để đưa ra được các tập tin có độ tương đồng cao nhất.

Chúng tôi so sánh tiêu đề dựa trên bản dịch của Google Translate API (Application Program Interface) từ tiếng Hoa sang tiếng Việt. Tuy việc sử dụng bản dịch này không hoàn toàn chính xác nhưng kết quả từ bản dịch của Google Translate khá là tốt nên chúng tôi sử dụng để so sánh độ tương đồng cũng mang lại kết quả tương đối. Các tiêu chí ngày, tháng, năm hoặc số phần trăm chúng tôi so sánh dựa trên số lượng tương đồng giữa các tập tin và không xét nội dung của các tập tin này. Ví dụ, tập tin A có số lượng ngày, tháng, năm là 3(20/10/1992, 5/1993, 5/6/1993) thì chúng tôi tìm các tập tin có số lượng ngày, tháng, năm cũng tương đồng như trên, hoặc ít hơn. Từ đó chúng tôi đưa ra được các tập tin có số phần trăm tương đồng theo ngày, tháng, năm. Đối với so sánh tập số phần trăm, chúng tôi cũng sử dụng phương pháp so sánh tương tự như ngày, tháng, năm.

Để làm giảm không gian so sánh của tập dữ liệu, chúng tôi còn dựa vào dung lượng của tập tin để so sánh. Đối với phương pháp này, trước khi đọc nội dung bên trong tập tin chúng tôi kiểm tra dung lượng của tập tin đó trước. Đối với tập tin văn bản sự chênh lệch 10kb là rất lớn. Nếu hai bài báo của tiếng Hoa và tiếng Việt đang cùng viết về một nội dung, nhưng hai tập tin có dung lượng nhiều hơn hoặc ít hơn lên đến 10kb hoặc lớn hơn, thì hai tập tin đó hầu như không cùng chứa một nội dung.

Chúng tôi đã tiến hành thử nghiệm trên một số thể loại có sự tương đồng về số lượng tập tin giữa hai ngôn ngữ của trang báo https://www.vietnamplus.vn/, cụ thể chúng tôi sẽ trình bày trong phần 4(Thử nghiệm). Do không có nhiều thời gian nên chúng tôi chỉ thử nghiệm trên một số tập dữ liệu có độ tương đồng cao về số lượng tập tin, để đảm báo số bài báo bên tiếng Hoa không quá ít so với số bài báo bên tiếng Việt.

Phần còn lại của bài báo này theo cấu trúc như sau: phần 2 trình bày một số công trình liên quan, phần 3 trình bày chi tiết về các phương pháp tìm kiếm tập tin song ngữ, phần 4 trình bày về các dữ liệu chúng tôi đã thử nghiệm. Cuối cùng, phần 5 là lời kết**.**

1. Công trình liên quan

Trong phần này chúng tôi, chúng tôi tập trung vào việc lọc các tập tin có cùng nội dung từ tập dữ liệu mà chúng tôi đã thu thập dựa trên các trang báo mạng song ngữ. Để tìm hiểu thêm về các phương pháp lọc các tập tin từ tập dữ liệu không thực sự song song.

Trong tài liệu tham khảo [2] Pascale Fung và Percy Cheung nghiên cứu về phương pháp khai thác các cặp câu song ngữ từ kho dữ liệu ban đầu rất thực sự không song song dựa trên phương pháp *via Bootstrapping* và *EM*, với nghiên cứu này có bước xử lý, tìm lọc các tập tin có cùng nội dung với nhau. Họ sử dụng phương pháp Bootstrapping lặp đi lặp lại dựa trên nguyên tắc “find-one-get-more” (cho phép thêm nhiều câu song ngữ từ tập dữ liệu không song ngữ). Với nghiên cứu này, cho ra kết quả tốt, phù hợp với tập dữ liệu của chúng tôi (dữ liệu rất không song song).

Trong tài liệu tham khảo [4] nghiên cứu về phương pháp lấy dữ liệu từ báo mạng cùng một nội dung nhưng trên nhiều ngôn ngữ khác nhau dựa trên *Mô hình Vector không gian* trên các thuật ngữ mới để tính toán. Với bài báo này, T.Pattabhi R.K Rao and Sobha Lalitha Devi đã đưa ra một phương pháp thu thập dữ liệu, mà không cần hậu xử lý mà vẫn có được tập dữ liệu song ngữ hoàn hảo.

Với nghiên cứu [4] đưa cho chúng tôi một cách nhìn mới về thu thập dữ liệu, mà không cần phải dùng các phương pháp lọc các tập tin sau khi thu thập. Chúng tôi có thể hiện thực lại phương pháp này nếu tập dữ liệu của chúng tôi sau khi xử lý không mang lại hiệu quả cao.

1. Phương pháp tìm kiếm tập dữ liệu song ngữ

Phương pháp của chúng tôi được mô phỏng trong hình 1, từ tập dữ liệu ban đầu, sau khi đã phân loại chúng tôi xử lý theo các bước sau:

1. Kiểm tra dung lượng tập tin trước khi so sánh.
2. So sánh tiêu đề của các tập tin với nhau.
3. So sánh danh sách ngày, tháng, năm(nếu có).
4. So sánh danh sách phần trăm (nếu có)
5. Tổ hợp các tiêu chí so sánh.

Để thuận lợi cho việc so sánh tiêu đề giữa hai ngôn ngữ Hoa-Việt chúng tôi đã dịch tiêu đề của các bài báo tiếng Hoa dựa trên bộ dịch Google Translate. Bộ dịch này đưa ra kết quả chính xác khoảng 70 🡪 80 %. Nhờ vào bộ dịch của Google chúng tôi có thể so sánh tiêu đề của tiếng Việt với tiêu đề của tiếng Hoa, từ đó đưa ra số phần trăm tương đồng giữa các tiêu đề với nhau.

Với tiêu chí so sánh tiêu đề của bài báo không đủ để kết luận các tập tin là tương đồng với nhau ở nội dung, do một số cách đặt tiêu đề không thống nhất, đặc biệt là tiếng Hoa ở các trang báo hay sử dụng tiêu đề như “越通社简讯 2017.6.22” tạm dịch là “Bản tin tiếng Việt ngày 22/6/2017”. Có rất nhiều bài báo với tiêu đề như ví dụ ở trên chỉ thay đổi thời gian, đối với tiếng Việt không có bất kỳ tiêu đề nào “Bản tin tiếng Việt”, do nội dung tiêu đề này không phù hợp với các bài báo dành cho người Việt đọc. Do đó, chúng tôi rút trích thêm các tiêu chí ngày, tháng, năm và số phần trăm bên trong nội dung của bài báo.

Để so sánh tiêu đề giữa hai tập tin với nhau chúng tôi sử dụng phương pháp so sánh dựa vào thuật toán *Levenshtein distance*. Với giải thuật này là thước đo sự khác biệt giữa hai chuỗi. Cụ thể đầu ra của giải thuật cho ra số lượng tối thiểu của ký tự được sửa đổi như chèn, xóa, thay thế cần thiết để thay đổi một từ vào một từ khác.

Với lev là khoảng cách giữa hai chuỗi a và b(chiều dài ) được tính bằng công thức . Khi dựa trên công thức thì khoảng cách bằng 0 khi và bằng 1 khi và là khoảng cách giữa ký tự i đầu tiền của chuỗi a với ký tự j đầu tiên của chuỗi b. Công thức của thuật toán như sau:

Ví dụ, dựa vào công thức trên thì khoảng cách giữa *flaw* và *lawn* là 2, vì phải thay đổi ký tự 2 lần để từ *flaw* thành từ *lawn* và không có cách thay đổi nào khác ít hơn 2 lần thay đổi.

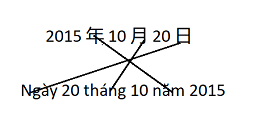
Khi chúng tôi hiện thực thuật toán *Levenshtein distance* chúng tôi chuyển các ký tự về dạng viết thường, để đảm bảo so sánh có hiệu quả và tối ưu nhất. Sau khi tìm được khoảng cách chúng tôi so sánh độ tương đồng giữa hai tiêu đề dựa trên công thức sau :

Với công thức so sánh độ tương đồng giữa hai chuỗi thì **ps(a,b)** là số phần trăm tương đồng giữa hai chuỗi a và b. Còn ,lần lượtlà chiều dài của chuỗi a và b. Nếu lớn hơn thì  **=** vàvà ngược lại. Còn là hàm tính khoảng cách giữa dựa trên thuật toán *Levenshtein distance* và ***l =*** .

Về ngày, tháng, năm của hai ngôn ngữ có sự khác nhau về vị trí như hình 1 mô tả vị trí ngày, tháng, năm của hai ngôn ngữ Hoa-Việt. Do đó, chúng tôi phải đưa ra các cách rút trích khác nhau. Khi rút trích ngày, tháng, năm chúng tôi hiện thực các bước theo bảng 1 phía dưới:

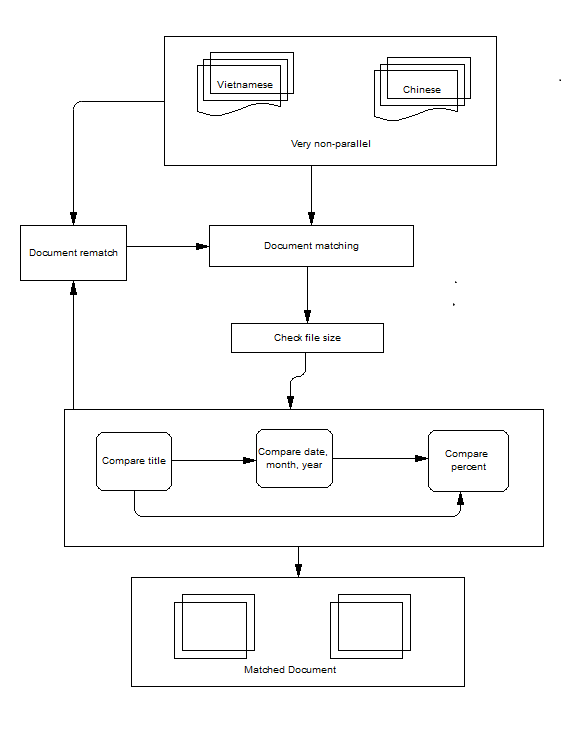
|  |
| --- |
| Rút trích ngày, tháng, năm của tiếng Hoa |
| 1: Kết nối với tập tin, đọc toàn bộ nội dung của tập tin. |
| 2: Đánh dấu vị trí có chứa từ khóa (日,月,年) trong nội dung . |
| 3: Rút trích ngày, tháng, năm hoặc tháng, năm từ các vị trí đã đánh dấu(lấy các ký tự phía trước của từ khóa tùy theo từ khóa là ngày hay tháng). |
| 4: Chuyển qua định dạng dd//mm/yyyy hoặc mm/yyyy và đưa vào danh sách ngày, tháng, năm của tập tin. |
| Rút trích số phầm trăm bên trong nội dung |
| 1: Kết nối với tập tin, đọc toàn bộ nội dung của tập tin. |
| 2: Đánh dấu vị trí có chứa từ khóa (%) trong nội dung . |
| 3: Rút trích phần trăm từ vị trí đã đánh dấu (lấy cái ký tự phía trước từ khóa) |
| 4: Đưa vào danh sách phần trăm của tập tin |

Bảng 1: Các bước rút trích ngày, tháng, năm và phần trăm



Hình 1: Vị trí ngày, tháng, năm của hai ngôn ngữ Hoa-Việt

Đối với ngày, tháng, năm của tiếng việt chúng tôi cũng rút trích tương tự như cách trên. Chỉ khác từ khóa và lấy các ký tự phía sau của từ khóa. Sau khi rút trích ngày, tháng, năm của cả tiếng Việt và tiếng Hoa, chúng tôi tiến hành so sánh ứng với mỗi tập tin tiếng Hoa với danh sách ngày, tháng, năm bên trong tập tin so sánh với các tập tin bên tiếng Việt. Kết quả sẽ thu được, ứng với mỗi tập tin tiếng Việt sẽ cho ra các danh sách tập tin có trùng ngày, tháng, năm và số phần trăm trùng lặp.



Hình 1: Sơ đồ biểu diễn phương thức tìm các tập tin tương đồng.

Để kết luận hai tập tin tương đồng bao nhiêu phần trăm chúng tôi dự trên công thức sau đây:

Về cơ bản, chúng tôi ký hiệu Q là số phần trăm tương đồng, với A là tập tin tiếng Việt, B là tập tin tiếng Hoa, với n là số lượng đối tượng ngày, tháng, năm tương đồng giữa hai tập tin A và B, là số lượng đối tượng ngày, tháng, năm trong tập tin A và B. Mỗi đối tượng ngày, tháng, năm là một ngày, tháng, năm cụ thể mà chúng tôi đã rút trích được từ nội dung.

Đối với cách so sánh các con số phần trăm từ nội dung (nếu có). Chúng tôi cũng hiện thực theo công thức trên.

1. Thử nghiệm
   1. Dữ liệu thử nghiệm

Chúng tôi sử dụng tập dữ liệu mà chúng tôi đã thu thập. Trong đó, chúng tôi chọn loại nội dung Chính trị của trang báo https://www.vietnamplus.vn/. Do nội dung này có lượng dữ liệu nhiều nhất, các loại nội dung khác chủ yếu số lượng bài báo tiếng Hoa ít hơn nhiều lần số liệu bài báo tiếng Việt, cụ thể có 12,808 tập tin tiếng Việt và 15,287 tập tin tiếng Hoa.

## Kết quả thử nghiệm

Chúng tôi thử nghiệm trên môi trường Java, trên hệ điều hành Windows 10 64 bits với một số thông tin cấu hình như 12GB RAM, Intel core i7-4510U CPU, NVIDIA GeForce 840M. Tốc độ kết nối mạng ổn định với tốc độ download là 4.5MB/s và tốc độ upload là 4.6 MB/s.

Với tiêu chí so sánh tiêu đề, đầu tiên chúng tôi chúng tôi dựa vào Google Translate API để dịch sau đó so sánh với tất cả các tiêu đề của tập tin trong tập dữ liệu báo tiếng Việt. Với mỗi tập tin, chúng tôi tìm bản dịch có độ tương đồng được tính theo phần trăm lớn nhất. Sau thử nghiệm, kết quả đạt được từ 33% đến 97%. Phần lớn các tập tin không tìm được bản dịch tiêu đề hoàn hảo. Hầu hết các tập tin nằm trong khoảng 30%-70% chiếm khoảng 95% trên tổng số tập tin thử nghiệm. Thời gian để thử nghiệm hơn 26 tiếng với 15287 lần dịch tiêu đề và khoảng hơn 150 triệu lần so sánh tiêu đề giữa các tập tin với nhau (sau khi đã loại bỏ các tập tin có dung lượng khác nhau quá lớn). Kết quả thống kê cụ thể cụ thể ở bảng 2, do có nhiều kết quả từ 33% 🡪 97% nên chúng tôi thống kê theo khoảng.

Theo chúng tôi nhận định, nguyên nhân làm cho kết quả sử dụng bản dịch của Google Translate không đạt được kết quả tốt vì tập dữ liệu ban đầu. Chúng tôi thu thập tập dữ liệu ban đầu từ báo mạng, hầu hết do các đội phát triển khác nhau viết. Có thể các đội viết báo muốn diễn đạt cùng nội dung nhưng họ lại sử dụng tiêu đề khác nhau để đặt cho bài báo đó.

|  |  |
| --- | --- |
| Khoảng phần trăm | Số lượng cặp tập tin |
| 33%-40% | 1752 |
| 41%-50% | 4989 |
| 51%-60% | 3979 |
| 61%-70% | 1406 |
| 71%-80% | 529 |
| 81%-90% | 139 |
| 90%-97% | 14 |

Bảng 2: Kết quả thử nghiệm so sánh tiêu đề dịch tiếng Hoa qua tiếng Việt.

Chúng tôi tiếp tục sử dụng tiêu chí rút trích ngày, tháng, năm (nếu có) để so sánh về nội dung bên trong. Kết quả thử nghiệm cụ thể trong bảng 3 bên dưới . Nhìn vào kết quả có 462 tập tin tương đồng về ngày, tháng, năm là 100%, chiếm hơn 38% trên tổng số lượng tập tin có nội dung chứa ngày, tháng, năm. Nhưng có rất nhiều tập tin chỉ chứa một đối tượng ngày, tháng, năm hoặc tháng, năm duy nhất.

|  |  |  |  |
| --- | --- | --- | --- |
| Phần trăm tương đồng | Số lượng cặp tập tin | | |
| 8.333334 % | 4 | | |
| 9.090909 % | 2 | | |
| 10.0 % | 14 | | |
| 11.111112 % | 5 | | |
| 12.5 % | 6 | | |
| 14.285715 % | 12 | | |
| 16.666668 % | 48 | | |
| 20.0 % | 50 | | |
| 22.222223 % | 3 | | |
| 25.0 % | 70 | | |
| 28.57143 % | 1 | | |
| 33.333336 % | 132 | | |
| 37.5 % | 3 | | |
| 40.0 % | 8 | | |
| 42.857143 % | | 4 |
| 50.0% | | 326 |
| 66.66667 % | | 41 |
| 75.0 % | | 7 |
| 85.71429 % | | 1 |
| 100.0 % | | 462 |
| Tổng | | 1199 |

Bảng 3: Kết quả thử nghiệm so sánh ngày, tháng, năm.

Biểu đồ 1: Biểu đồ thể hiện kết quả thử nghiệm so sánh ngày, tháng, năm theo khoảng.

Chúng tôi tiếp tục thử nghiệm thêm phương pháp so sánh về phần trăm. Với thử nghiệm này, chúng tôi rút trích toàn bộ các con số phần trăm có trong nội dung, sau đó cho vào danh sách, rồi so sánh từng cặp tập tin với nhau. Với cách so sánh này, chúng tôi không lấy ra cặp tập tin có số tương đồng cao nhất, mà một tập tin có thể có một hoặc nhiều tập tin có cùng số phần trăm giống nhau. Chúng tôi thu được **2431** tập tin tiếng Việt và **2535** tập tin tiếng Hoa có chứa con số phần trăm kết quả. Kết quả so sánh của các tập tin này chúng tôi thống kê ở bảng 4 và biểu diễn trên biểu đồ 2 bên dưới.

|  |  |
| --- | --- |
| Phần trăm tương đồng | Số lượng cặp tập tin |
| 8%-20% | 78795 |
| 21%-30% | 35202 |
| 31%-40% | 38800 |
| 41%-50% | 39682 |
| 51%-60% | 1217 |
| 61%-70% | 4249 |
| 71%-80% | 1050 |
| 81%-90% | 266 |
| 91-100% | 29340 |

Bảng 4: Kết quả thử nghiệm so sánh phần trăm của các tập tin.

Biểu đồ 2: Biểu đồ thể hiện kết quả so sánh theo tiêu chí phần trăm.

Nhìn vào kết quả biểu đồ 2 cho thấy, các cặp tin tương đồng với nhau với độ tương đồng không hoàn toàn chiếm khá nhiều, chủ yếu các lần so sánh, cho ra kết quả dưới 50% tương đồng. Nhưng ở kết quả 100% độ tương đồng chúng tôi cũng thu được một số lượng đáng kể. Với các tập tin này, chúng tôi sẽ kết hợp thêm với các tiêu chí đã so sánh, để tìm ra các tập tin có nội dung tương đồng tốt nhất.

1. Kết luận

Hiện tại chúng tôi đã thử nghiệm trên 4 phương pháp để tìm kiếm các tập tin có cùng nội dung là : dựa trên dung lượng của tập tin để làm giảm không gian so sánh, so sánh dựa trên tiêu đề của bài báo, so sánh dựa trên nội dung có chứa ngày, tháng, năm và phần trăm bên trong nội dung. Đối với các phương pháp trên, chúng tôi đã cho ra được danh sách các tập tin có cùng nội dung với nhau với hiệu quả cao.

Chúng tôi sẽ thêm các tiêu chí để tìm ra các tập tin có cùng nội dung với nhau như tên người riêng và tên địa danh riêng để làm tăng số phần trăm tương đồng của các tập tin mà chúng tôi đã xác định là có cùng nội dung.

Với tập dữ liệu mà chúng tôi xác định dựa trên các phương pháp đã thử nghiệm, chúng tôi sẽ thử nghiệm thêm nhiều tập dữ liệu khác, để có nguồn dữ liệu dồi dào, phục vụ cho bước tiếp theo là tìm các cặp câu song ngữ Hoa-Việt từ tập dữ liệu này.

# **References**

[1] Thuy Vu, Ai Ti Aw, Min Zhang, *Feature-based Method for Document Alignment in Comparable News Corpora.*

[2] Pascale Fung and Percy Cheung, *Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM.*

[3] Marek Medved, Milos Jakubicek, Vojtech Kovar, *English-French Document Alignment Based on Keywords and Statistical Translation.*

[4] T.Pattabhi R.K Rao and Sobha Lalitha Devi, *How to Get the Same News from Different Language News Papers.*

[5] Christian Buck and Philipp Koehn, *Shared Task Bilingual Document Alignment.*

[6] Ulrich Germann, *Bilingual Document Alignment with Latent Semantic Indexing.*