

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG**

LÊ CÔNG VỖ

**NGHIÊN CỨU PHƯƠNG PHÁP THU THẬP TẬP
DỮ LIỆU SONG SONG (PARALLEL CORPUS)
VIỆT – ANH, VIỆT – PHÁP TỪ CÁC NGUỒN TÀI
LIỆU ĐA NGỮ**

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số: 60.48.01

LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2011

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: TS. Huỳnh Công Pháp

Phản biện 1: PGS. TS. Đoàn Văn Ban

Phản biện 1: TS. Huỳnh Hữu Hưng

Luận văn sẽ được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng ngày 10 tháng 9 năm 2011.

Có thể tìm hiểu luận văn tại:

- Trung tâm Thông tin – Học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng

MỞ ĐẦU

1. Lý do chọn đề tài

Trước sự bùng nổ của thông tin như hiện nay, nhu cầu dịch tự động ngày càng lớn. Việc dịch tự động các văn bản, bài báo, tạp chí,...(từ ngôn ngữ A sang ngôn ngữ B hoặc nhiều ngôn ngữ khác) để phục vụ công việc học tập, nghiên cứu một cách kịp thời, nhanh chóng là một trong những vấn đề hết sức cấp bách. Chính vì vậy, bài toán dịch tự động song song giữa các ngôn ngữ càng có ý nghĩa thiết thực hơn bao giờ hết.

Ngày nay, với sự phát triển của CNTT, Internet phát triển mạnh và được sử dụng rộng rãi. Các dịch vụ dịch thuật ngày càng trở nên phổ biến và không thể thiếu của con người thì việc xây dựng kho dữ liệu song song (Parallel Corpus) là điều cần làm để làm cơ sở, nền tảng cho hệ thống dịch tự động thông minh hơn nhằm phục vụ cho việc dịch các đoạn văn bản, các bài báo,...

Ngoài ra kho dữ liệu song song cũng rất quan trọng đối với mục đích nghiên cứu và học tập như là tạo hệ thống mẫu để đánh giá chất lượng dịch, phục vụ cho việc học ngoại ngữ,...

Ngày nay, việc đánh giá và phát triển các hệ thống ngôn ngữ khác như: Tiếng Anh, tiếng Ý, tiếng Pháp,... dễ dàng và đạt được chất lượng cao. Bởi vì, các ngôn ngữ này có hệ thống kho ngữ liệu song song lớn và có chất lượng rất cao. Nhưng đối với tiếng Việt rất khó khăn để xây dựng một hệ thống kho ngữ liệu song song chất lượng tốt và hiện nay chưa tồn tại dịch tự động có chất lượng cao. Ví dụ như GoogleTranslator có hỗ trợ tiếng Việt nhưng có chất lượng cũng chỉ khoảng 20% so với chất lượng của con người dịch mà thôi. Nguyên

nhân cơ bản là các hệ thống này vẫn còn thiếu tập dữ liệu dịch tự động song song có chất lượng tốt, đủ lớn.

Hiện nay, tồn tại nhiều phương pháp để thu thập kho ngữ liệu song song cho nhiều ngôn ngữ khác trên thế giới. Tuy nhiên, đối với tiếng Việt gặp nhiều khó khăn trong việc xây dựng kho ngữ liệu song song. Ví dụ như: Tính nhập nhằng của tiếng Việt, sự phức tạp của từ nhiều âm tiết,... Như vậy, chúng tôi có thể áp dụng các phương pháp có sẵn và cải tiến để thu thập được kho ngữ liệu song song cho tiếng Việt. Đó là chọn lựa phương pháp thu thập từ nguồn tài nguyên sẵn có: Từ các Website đa ngữ trên Internet. Bởi vì, Internet có rất nhiều các Website đa ngữ chứa tiếng Việt, tiếng Anh, tiếng Pháp,... Và trong thời gian giới hạn đề tài, chúng tôi chỉ chọn các cặp ngôn ngữ Việt – Anh, Việt – Pháp vì đây là hai ngôn ngữ chủ đạo trong cộng đồng người Việt và tính cấp bách của nó đối với người Việt. Trong tương lai, chúng tôi sẽ phát triển cho các cặp ngôn ngữ khác.

2. Mục đích nghiên cứu

Mục tiêu mà đề tài hướng đến là nghiên cứu các phương pháp trích lọc các câu song song Việt – Anh, Việt – Pháp từ các Website đa ngữ. Đồng thời nghiên cứu các phương pháp và mô hình cho phép con người cộng tác để cải tiến, sửa đổi,... kho ngữ liệu song song.

Thu thập được một kho ngữ liệu song song Việt – Anh, Việt – Pháp có chất lượng tốt từ các Website đa ngữ. Sau đó xây dựng mô hình quản lý và cho phép người dùng tương tác nhằm cải tiến hệ thống kho ngữ liệu song song cho việc dịch tự động.

3. Đối tượng và phạm vi nghiên cứu

Hiện nay, trên thế giới có rất nhiều kho ngữ liệu song song cho các cặp ngôn ngữ khác nhau. Tuy nhiên, trong khuôn khổ luận văn này chúng tôi chỉ nghiên cứu về phương pháp thu thập kho ngữ liệu song song Việt – Anh và Việt – Pháp và mô hình cho phép con người cộng tác để cải tiến, sửa đổi,... kho ngữ liệu song song đã thu thập được.

Ngày nay, Internet rất quan trọng và là một phần không thể thiếu đối với đời sống con người. Với sự hội nhập quốc tế các công ty, các cơ quan, các tổ chức,... đều có Website riêng để phục vụ cho việc giao tiếp và quảng bá thương hiệu. Để tiện giao tiếp và đưa thương hiệu của mình vươn tầm quốc tế, hầu hết đã xây dựng cho mình một Website đa ngữ để đáp ứng yêu cầu đó. Nên số lượng các Website đa ngữ là rất lớn trên Internet. Đó là một nguồn tài nguyên đa ngữ cực kỳ lớn chưa được khai thác triệt để. Trong luận văn này chúng tôi chọn nguồn dữ liệu này để thu thập và xây dựng kho ngữ liệu song song.

4. Phương pháp nghiên cứu

Để đáp ứng mục tiêu và nhiệm vụ của đề tài đặt ra, tôi áp dụng hai phương pháp nghiên cứu:

Phương pháp lý thuyết:

Trong phương pháp này tôi nghiên cứu cơ sở lý thuyết về kho ngữ liệu song song, phương pháp phân tích, trích lọc dữ liệu từ các Website đa ngữ, phương pháp phân đoạn, liên kết kết dữ liệu để xây dựng kho ngữ liệu song song.

Phương pháp thực nghiệm:

Xây dựng các module trích lọc, phân đoạn, liên kết giữa các cặp câu song song của các cặp ngôn ngữ. Đồng thời xây dựng chương

trình quản lý cho phép người dùng cộng tác, sửa đổi, cải tiến,... kho ngữ liệu song song ngày càng tốt hơn.

5. Ý nghĩa khoa học và thực tiễn của đề tài

Ý nghĩa khoa học:

Góp phần xây dựng hệ thống dịch tự động, nâng cao chất lượng hệ thống dịch tự động. Có thể dùng tập dữ liệu đó như một tập dữ liệu khoa học đáng tin cậy để những người muốn nghiên cứu và ứng dụng trong việc dịch tự động song ngữ có thể dùng nó để tham khảo và phát triển.

Ý nghĩa thực tiễn:

Xây dựng tập dữ liệu dịch tự động song ngữ phục vụ cho hệ thống dịch tự động đa ngữ song song và chương trình quản lý tập dữ liệu song song các ngôn ngữ.

Nó có ý nghĩa hết sức thiết thực trong việc dịch thuật tự động đa ngữ song song, kết quả của đề tài góp phần phục vụ tốt hơn cho việc dịch tự động nhiều ngôn ngữ. Ngoài ra chúng ta có thể sử dụng kho ngữ liệu song song này trong việc dạy và học ngoại ngữ trong nhà trường.

6. Bố cục của luận văn

Bố cục của luận văn gồm 3 chương, sau phần mở đầu là:

Chương 1 chúng tôi giới thiệu sơ bộ những nội dung tổng quan về kho ngữ liệu song song và phương pháp thu thập, trích lọc dữ liệu từ các Website đa ngữ. Trong phần này cũng trình bày tổng quan về đặc điểm của các Website đa ngữ và phương pháp xác định các Website đa ngữ.

Trong chương 2, luận văn giới thiệu về các bước xây dựng kho dữ liệu song ngữ và phương pháp quản lý kho ngữ liệu song song đã thu thập được, đi sâu vào việc xây dựng và đo độ tương đồng về ngôn ngữ của các câu đã thu thập được.

Trong chương này tập trung nghiên cứu phân tích xây dựng kiến trúc tổng thể của hệ thống gồm các thành phần liên quan, cách vận hành của hệ thống. Từ kiến trúc tổng thể đã xây dựng, chúng tôi khai thiết kế các thành phần đã phân tích, xây dựng cơ sở dữ liệu, ứng dụng quản lý kho ngữ liệu song song đã được xây dựng.

Phần kết luận, tổng hợp những kết quả nghiên cứu chính của luận văn, chỉ ra một số hạn chế chưa hoàn thiện. Đồng thời, luận văn cũng đề xuất một số hướng nghiên cứu cụ thể tiếp theo của tác giả.

CHƯƠNG 1 – CƠ SỞ LÝ THUYẾT

Trong chương này, chúng tôi đã trình bày khái quát về cơ sở lý thuyết về: Hệ thống dịch tự động, một số khái niệm về kho ngữ liệu song song, tổng quan về một số kho ngữ liệu song song và một số bước quan trọng trong qui trình xây dựng kho ngữ liệu song ngữ hiện nay.

1.1. HỆ THỐNG DỊCH TỰ ĐỘNG

1.1.1. Các hệ thống dịch tự động

1.1.2. Sự cần thiết của việc xây dựng các hệ thống dịch tự

động có hỗ trợ tiếng Việt

Dịch tự động là dạng đặc biệt trong hoạt động dịch thuật của biên phiên dịch viên. Đồng thời, dịch tự động là phương tiện đặc biệt trong

giao tiếp giữa các ngôn ngữ và cung cấp thông tin kinh tế, xã hội hiện đại cần thiết không những chỉ đối với các quốc gia trên toàn thế giới, mà còn cần thiết đối với cá nhân mỗi con người trong hoạt động đời sống xã hội của mình.

Là một nước đang phát triển, Việt Nam cần sử dụng nhiều công nghệ tiên tiến của thế giới. Trong số các công nghệ tiên tiến đang được các nước trên thế giới sử dụng nhiều nhất hiện nay có công nghệ dịch tự động. Việc cần thiết sử dụng các hệ thống dịch tự động tiếng Việt Nam bắt nguồn từ hàng loạt các nguyên nhân sau đây:

Thứ nhất: Ở Việt Nam hiện nay, tăng mạnh nhu cầu đọc nhanh để lấy được nhiều thông tin mới, bổ ích, đặc biệt là các thông tin về khoa học, công nghệ và kỹ thuật mới từ các nước tiên tiến trên thế giới.

Thứ hai: Trong tất cả các trường học và công sở của Việt Nam ngày càng đòi hỏi sự nắm vững tiếng Anh như một phương tiện cần thiết trong giao tiếp giữa các ngôn ngữ.

Thứ ba: Nhu cầu giao lưu bằng nhiều ngoại ngữ về chuyên ngành khoa học ngày càng tăng.

Thứ tư: Sự phát triển của công nghệ máy tính và Internet đã thúc đẩy nhu cầu khai thác thông tin trên nhiều trang Web bằng nhiều ngôn ngữ khác nhau.

Cần thừa nhận rằng dịch tự động không thể thay thế hoàn toàn việc dịch thuật được thực hiện bởi con người. Tuy nhiên, nhờ hệ thống tự động trong chương trình dịch bằng máy, có thể nâng cao đáng kể hiệu quả lao động của người dịch và chất lượng của cán bộ chuyên ngành. Ngày nay, dịch tự động đã phát triển đến mức có thể thực hiện dịch trực tiếp trên mạng. Tuy vậy, trong số các ngôn ngữ có thể dịch tự động trực tuyến, chưa có tiếng Việt hoặc có nhưng chất lượng rất thấp.

Rõ ràng rằng trong các điều kiện hiện nay, khi người dịch chuyên nghiệp cần phải biết vô vàn các thuật ngữ của nhiều chuyên ngành và cần phải thuộc rất nhiều tên gọi chính xác các chủng loại chi tiết, linh kiện, dụng cụ, cơ cấu, các chất v.v. khác nhau, thì một phiên dịch viên dù giỏi đến đâu cũng không thể cập nhật hết được một lượng thông tin khổng lồ trong nền kinh tế toàn cầu phát triển. Lúc này, việc sử dụng dịch tự động để trợ giúp cho quá trình dịch thuật là tất yếu và cần thiết.

1.2. MỘT SỐ KHÁI NIỆM VỀ KHO NGỮ LIỆU

1.2.1. Kho ngữ liệu (Corpus)

1.2.2. Kho ngữ liệu đa ngữ (Multilingual Corpora)

1.2.3. Kho ngữ liệu so sánh (Comparable Corpus)

1.2.4. Kho ngữ liệu song song (Parallel Corpus)

Kho ngữ liệu song song là một thành phần thiết yếu quan trọng không thể thiếu. Chất lượng của kho ngữ liệu song song đóng vai trò quyết định đến chất lượng đầu ra của hệ dịch. Hệ thống dịch sẽ không thể cho kết quả tốt nếu kho ngữ liệu song song sử dụng trong quá trình huấn luyện có chất lượng không tốt cho dù được áp dụng các phương pháp học máy tiên tiến nhất.

Kho ngữ liệu song song có thể được thu thập từ nhiều nguồn. Tổng quát, có thể chia các nguồn này thành hai loại, nguồn ngữ liệu ở dạng giấy viết và nguồn ngữ liệu dạng điện tử. Kho ngữ liệu song song ở dạng giấy viết có thể được tìm thấy dễ dàng trong các sách học ngoại ngữ, các sách truyện, tài liệu song ngữ và các từ điển song ngữ. Việc tìm kiếm loại ngữ liệu này đơn giản, tuy nhiên quá trình nhập liệu vào máy tính tốn nhiều thời gian và công sức. Do vậy trong thực tế, gần

như không có nhóm nghiên cứu nào thực hiện việc xây dựng kho ngữ liệu song song từ nguồn này.

Với phát triển bùng nổ của công nghệ thông tin và Internet, các kho ngữ liệu song song dưới dạng điện tử hiện đang tồn tại khá phong phú ở nhiều dạng. Kho ngữ liệu song song đơn giản nhất mà chúng ta có thể nhận thấy là các loại truyện song ngữ, các tài liệu và tác phẩm văn học nước ngoài nổi tiếng được dịch sang tiếng Việt. Nguồn ngữ liệu này có thể thu thập bằng tay từ Internet.

Kho ngữ liệu song song (Parallel Corpus) là một tập các văn bản (tài liệu) trong nhiều ngôn ngữ khác nhau, trong đó có một ngôn ngữ nguồn và một (hoặc nhiều) ngôn ngữ đích (được dịch từ ngôn ngữ nguồn).

1.2.5. Sự liên kết (Alignment)

1.3. TỔNG QUAN VỀ MỘT SỐ KHO NGỮ LIỆU SONG

NGỮ

1.3.1. Lịch sử hình thành và phát triển của kho ngữ liệu

song song

1.3.2. Một số kho ngữ liệu song song trên thế giới

1.3.3. Nội dung của các kho ngữ liệu

1.3.4. Cấu trúc của các kho ngữ liệu

1.3.5. Các công trình nghiên cứu trong nước về kho ngữ liệu

song song

1.4. TÌM HIỂU MỘT SỐ BƯỚC QUAN TRỌNG TRONG QUY TRÌNH XÂY DỰNG KHO NGŨ LIỆU SONG SONG

1.4.1. Xác định nguồn tài nguyên phục vụ việc xây dựng kho ngữ liệu song song

1.4.2. Liên kết dữ liệu đã thu thập

1.5. TỔNG KẾT CHƯƠNG 1

CHƯƠNG 2 – ĐỀ XUẤT PHƯƠNG PHÁP XÂY DỰNG VÀ QUẢN LÝ KHO NGŨ LIỆU SONG SONG

2.1. HIỆN TRẠNG NGHIÊN CỨU VỀ KHO NGŨ LIỆU SONG SONG

2.1.1. Tầm quan trọng của kho ngữ liệu song song

Hiện nay, các kho ngữ liệu song song đã và đang được xây dựng cho nhiều ngôn ngữ khác nhau trên thế giới nên tầm quan trọng và giá trị sử dụng của nó là rất lớn. Mặt khác, kho ngữ liệu song song là một trong những thành phần quan trọng nhất trong lĩnh vực xử lý ngôn ngữ tự nhiên và dịch tự động. Bởi vì, nhờ các kho ngữ liệu song song này chúng ta có thể xây dựng được các chương trình quan trọng trong các lĩnh vực này. Nghiên cứu cho thấy kho ngữ liệu song song có các đặc điểm rất quan trọng đối với khoa học cũng như đời sống của con người:

- Là thành phần không thể thiếu đối với hệ thống dịch tự động.

- Có thể sử dụng kho ngữ liệu song song để làm mẫu đánh giá một số hệ thống, chương trình tự động: hệ thống dịch tự động, chương trình học ngoại ngữ,...
- Phục vụ cho mục đích nghiên cứu về ngôn ngữ học, xử lý ngôn ngữ,... của các nhà nghiên cứu.
- Trong lĩnh vực nghiên cứu và giảng dạy ngôn ngữ trong nhà trường hiện nay.

Khi Việt Nam đã gia nhập WTO nên sự cần thiết phải thực hiện giao lưu để trao đổi thông tin ở mức độ giao tiếp bằng nhiều ngôn ngữ trên phạm vi toàn thế giới. Điều cần thiết trong việc hội nhập kinh tế thế giới là cần có hệ thống kho ngữ liệu song song quốc gia. Nhằm phục vụ các lĩnh vực liên quan đến nghiên cứu khoa học, giảng dạy, cập nhật thông tin trong và ngoài nước v.v. Mặt khác người phiên dịch chuyên nghiệp cần phải biết vô vàn các thuật ngữ của nhiều chuyên ngành và cần phải thuộc rất nhiều tên gọi chính xác các chủng loại chi tiết, linh kiện, dụng cụ, cơ cấu, các chất v.v. khác nhau. Thì một phiên dịch viên dù giỏi đến đâu cũng không thể cập nhật hết được một lượng thông tin khổng lồ đó. Lúc này, việc sử dụng kho ngữ liệu song song để trợ giúp cho quá trình dịch thuật là tất yếu và cần thiết.

2.1.2. Hiện trạng nghiên cứu về kho ngữ liệu song song hiện nay

2.2. PHƯƠNG PHÁP TIẾP CẬN

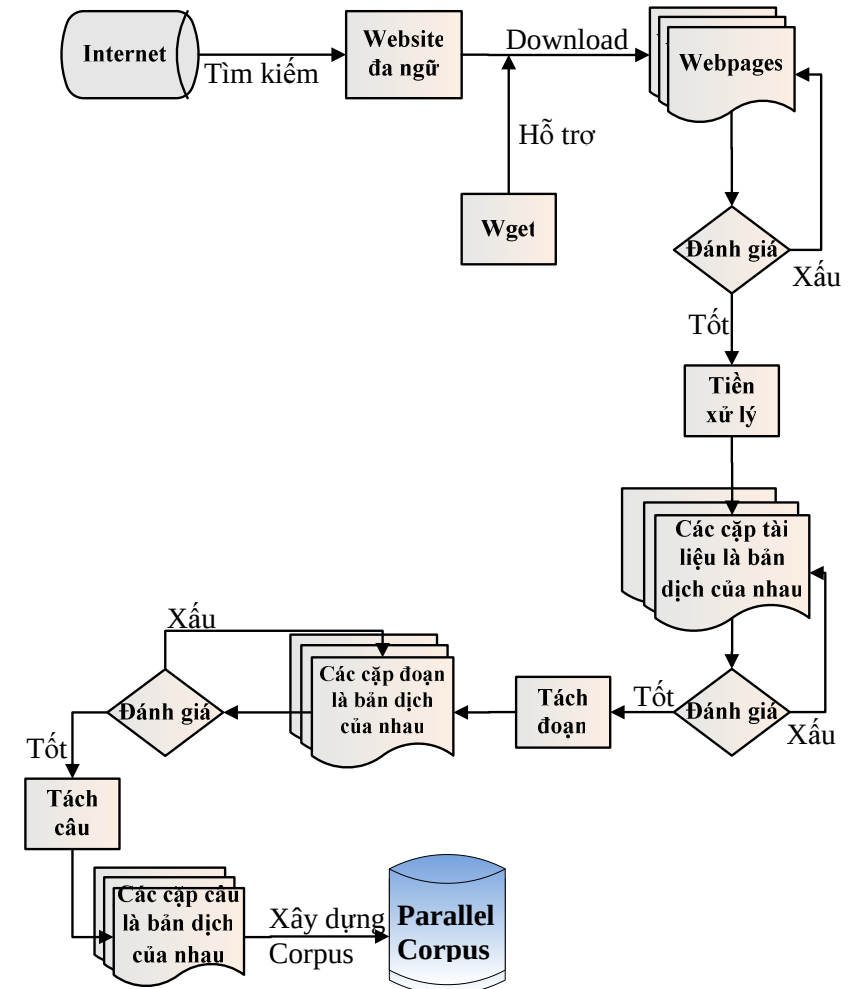
Với vai trò và tầm quan trọng của kho ngữ liệu song song đối với các ứng dụng xử lý ngôn ngữ tự nhiên, đồng thời được thúc đẩy bởi việc thiếu cơ sở dữ liệu song song Việt – Anh, Việt – Pháp cho nhiều nghiên cứu khác, luận văn tập trung vào các công việc:

- Tìm hiểu, nghiên cứu, phát triển các công nghệ trong bài toán trích lọc dữ liệu các câu song song từ các Website đa ngữ cho các cặp ngôn ngữ Việt – Anh, Việt – Pháp.
- Xây dựng công cụ trích lọc các cặp câu song song trên các Website đa ngữ cho các cặp ngôn ngữ Việt – Anh, Việt – Pháp.
- Có thể chia làm hai tiếp cận chính là tiếp cận dựa trên nội dung và tiếp cận dựa trên cấu trúc của trang Web. Đối với tiếp cận dựa trên nội dung, chúng ta phải sử dụng từ điển song ngữ. Do việc từ điển song ngữ Việt – Anh, Việt – Pháp có quá nhiều nhập nhằng, hơn nữa do thời gian có hạn nên chúng tôi tập trung vào nghiên cứu theo tiếp cận thứ hai là dựa vào cấu trúc văn bản (cấu trúc trang Web). Phương pháp được chúng tôi sử dụng và phát triển dựa trên nghiên cứu [3,5], với hai phần:
- Xác định các thuộc tính dùng để đo độ tương tự giữa hai trang HTML
- Áp dụng thuật toán học máy để xây dựng mô hình trên tập các thuộc tính trên.

Đối với phần xác định các thuộc tính để đo độ tương đồng giữa hai trang HTML, chúng tôi sẽ sử dụng các thuộc tính sau:

- So sánh độ tương đồng tên file của trang Web
- So sánh độ tương đồng cấu trúc URL
- So sánh cấu trúc HTML của cặp trang Web
- Và một số tiêu chí khác để làm giảm thời gian chạy của hệ thống như ngày sửa, ngày tạo, tỉ lệ âm tiết, tỉ lệ chunk.

Quá trình khai thác dữ liệu từ Internet để xây dựng hệ thống Kho ngữ liệu song song được thể hiện qua sơ đồ hình 2.1.



Hình 2.1. Quá trình khai thác dữ liệu và xây dựng kho ngữ liệu song song

Vậy quá trình này gồm các bước như sau:

Bước 1: Tìm kiếm và phát hiện các Website đa ngữ Việt – Anh, Việt – Pháp, Việt – Anh – Pháp và ngược lại; chúng tôi dựa vào hai tiêu chí:

- Dựa vào nội dung liên kết trong trang
- Dựa vào URL của trang

Bước 2: Download dữ liệu từ các trang Web đa ngữ đã tìm được ở bước 1 qua sự hỗ trợ của công cụ download: Wget, Teleport pro,... Việc dùng phần mềm nào để download còn tùy vào đặc điểm của từng trang.

Bước 3: Xác định các trang là bản dịch của nhau, đánh giá chất lượng bản dịch của các cặp trang Web và tiền xử lý các thẻ HTML, khôi phục các ký tự tiếng Việt, tiếng Pháp chúng ta sẽ được các cặp là bản dịch của nhau.

Bước 4: Đánh giá và tiến hành tách đoạn các cặp bản dịch ở bước 3, chúng ta thu được là tập dữ liệu các cặp dịch của các đoạn.

Bước 5: Đánh giá, tách câu và liên kết câu của các cặp đoạn dịch ở bước 4. Chúng ta thu được là tập các cặp câu dịch. Đây chính kho ngữ liệu song song.

Các bước trên sẽ được chúng tôi làm rõ ở các phần sau của luận văn.

2.3. PHÁT HIỆN CÁC WEBSITE ĐA NGỮ VIỆT – ANH, VIỆT – PHÁP

Phần này trình bày chúng tôi một số phương pháp phát hiện tự động các trang Web có tài liệu song ngữ Việt – Anh, Việt – Pháp. Do

dữ liệu trên Internet rất lớn, chúng ta không thể dò tìm từng trang một. Ý tưởng của chúng tôi là sử dụng một máy tìm kiếm (Search Engine) để tìm nhanh các trang Web có khả năng chứa tài liệu song ngữ dựa vào một số heuristic quan sát bằng mắt. Chúng tôi sử dụng các heuristic trước trình bày bên dưới. Trong luận văn này chúng tôi đưa ra một số phương pháp để xác định các trang Web đa ngữ.

2.3.1. Dựa vào nội dung liên kết (link) trong các Website

2.3.1.1. Website tiếng Anh có liên kết đến trang tiếng Việt

2.3.1.2. Website tiếng Pháp có liên kết đến trang tiếng Việt

2.3.1.3. Website tiếng Việt có liên kết đến trang tiếng Anh hoặc

tiếng Pháp

2.3.1.4. Website tiếng Việt có liên kết đến trang tiếng Anh và

tiếng Pháp

2.3.2. Dựa vào URL của trang

2.4. BÀI TOÁN KHAI THÁC DỮ LIỆU SONG SONG TỪ

CÁC WEBSITE ĐA NGỮ

2.4.1. Lọc theo cấu trúc

2.4.2. Lọc theo nội dung

2.4.3. Lọc theo các đặc điểm khác

2.4.4. Thuật toán qui hoạch động

2.5. XÂY DỰNG KHO DỮ LIỆU SONG SONG

Hiện nay, kho ngữ liệu song song ngày càng đóng vai trò quan trọng trong nhiều lĩnh vực như tìm kiếm xuyên ngữ (Cross-language Information Retrieval), dịch máy (Machine Translation),... Tuy nhiên, những kho ngữ liệu song song này hiện chỉ có cho một số cặp ngôn ngữ như Anh – Pháp, Anh – Hoa, do việc xây dựng chúng đòi hỏi rất nhiều thời gian và công sức. Hiện chúng ta chưa có kho ngữ liệu song song Việt – Anh, Việt – Pháp nào được công bố để phục vụ cho các nghiên cứu liên quan.

Do việc xây dựng kho ngữ liệu thủ công quá tốn kém, các nhà nghiên cứu bắt đầu tìm kiếm các phương pháp xây dựng tự động: Khai thác từ Internet. Số lượng các trang Web song ngữ Việt – Anh, Việt – Pháp trên Internet ngày càng nhiều do các cơ quan, tổ chức trong nước có khuynh hướng tự giới thiệu mình, hợp tác với nước ngoài. Vì vậy, Internet trở thành một nguồn cung cấp tài liệu song ngữ rất tiềm năng và đã có nhiều công trình nghiên cứu về lĩnh vực này.

Tuy nhiên, lượng dữ liệu lớn mà Internet có khả năng cung cấp cũng mang lại nhiều khó khăn. Do lượng dữ liệu quá lớn, việc tự động dò tìm các trang Web chứa tài liệu song ngữ là không dễ dàng. Ngay khi đã có được trang Web song ngữ, việc xác định những trang nào là dịch của nhau cũng không đơn giản do nó đòi hỏi nhiều tài nguyên về ngôn ngữ trong khi những tài nguyên hỗ trợ tiếng Việt còn rất hạn chế. Một khó khăn nữa là chất lượng tài liệu dịch trên Internet chưa phải là bản dịch chuẩn.

Do không được kiểm soát, chất lượng dịch của tài liệu trên Internet là rất khác nhau, và không phải tài liệu nào cũng được dịch chuẩn. Hơn nữa, có nhiều tài liệu (ví dụ tài liệu tin tức song ngữ Việt –

Anh, Việt – Pháp) cùng viết về một vấn đề (tin tức), nhưng lại dùng ngôn từ khác nhau, tổ chức ý khác nhau và thậm chí đưa ra những dẫn chứng khác nhau làm cho việc phát hiện những tài liệu thực sự song ngữ (chất lượng dịch tốt) càng trở nên khó khăn hơn.

Từ những vấn đề nêu trên, chúng tôi đã tham khảo các phương pháp xây dựng tự động kho ngữ liệu song song từ Internet cho các ngôn ngữ khác và phát triển phương pháp thu thập kho ngữ liệu song song Việt – Anh, Việt – Pháp tự động và bán tự động để phục vụ cho nghiên cứu trong nước. Chúng tôi đã cài đặt phương pháp của mình thành các Module với các chức năng sau:

- Tự động và bán tự động phát hiện các địa chỉ Web có tài liệu song ngữ Việt – Anh, Việt – Pháp.
- Tự động tải các Website về máy trạm để xử lý.
- Liên kết ở mức văn bản: Tải các trang Web từ các địa chỉ này về, tiền xử lý (loại bỏ thẻ HTML, các phần nội dung phụ,...) và tiến hành dò tự động để tìm các cặp tài liệu là dịch của nhau.
- Liên kết ở mức câu: Với mỗi cặp tài liệu dịch tìm thấy, tiến hành liên kết tự động, để có được kho ngữ liệu song song gồm các cặp câu tương ứng là dịch của nhau.
- Phần mềm quản lý kho ngữ liệu song song: Cho phép con người duyệt qua các cặp câu song song và chỉnh sửa, thêm bớt nếu cần.

2.5.1. Liên kết ở mức tài liệu: Phát hiện các cặp tài liệu song

ngữ

2.5.1.1. Tải tài liệu

2.5.1.2. Tiền xử lý

2.5.1.3. Tìm các cặp tài liệu là dịch của nhau

2.5.2. Xây dựng kho ngữ liệu song song

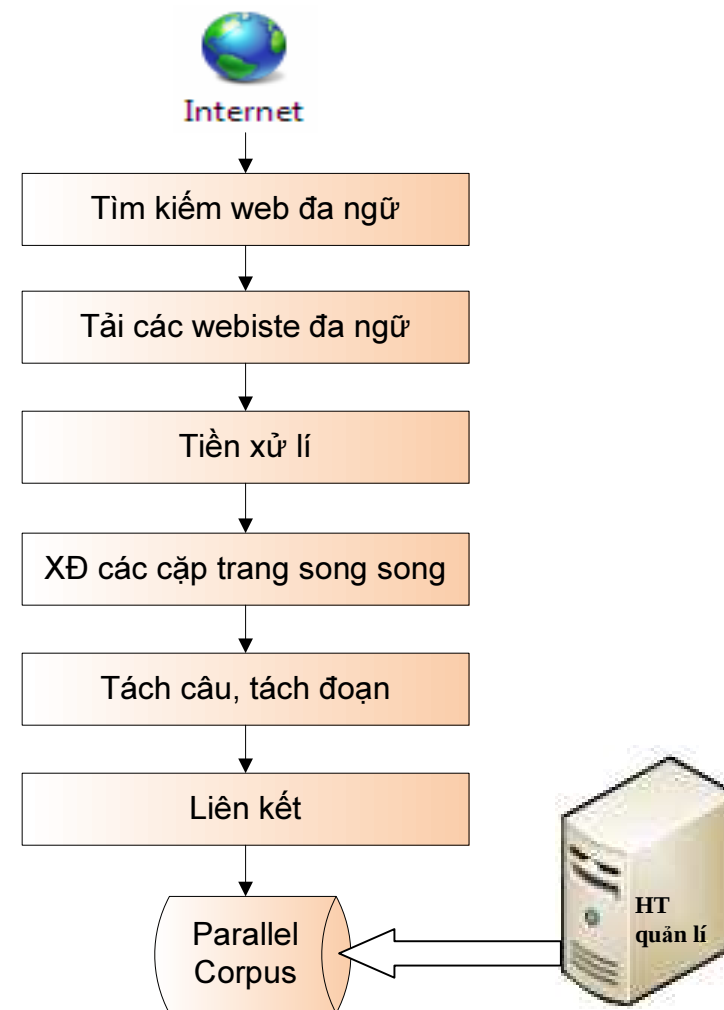
2.6. TỔNG KẾT CHƯƠNG 2

CHƯƠNG 3 – XÂY DỰNG VÀ PHÁT TRIỂN ỨNG DỤNG

3.1. XÂY DỰNG CÁC MODULE THÀNH PHẦN

3.1.1. Kiến trúc hệ thống

Trong chương 2 chúng tôi đã đề xuất phương pháp xây dựng và quản lý kho ngữ liệu song song. Trong chương này chúng tôi sẽ xây dựng các Module thành phần. Hệ thống được thiết kế như Hình 3.1



Hình 3.1. Quy trình xây dựng các Module

Đầu tiên hệ thống sẽ tìm kiếm các Website đa ngữ Việt – Anh, Việt – Pháp từ trên Internet, sau bước này chúng tôi sẽ được URL của các trang web. Sau đó tải dữ liệu từ các trang đã tìm được về máy tính.

Bước tiếp theo, chúng tôi tiến hành tiền xử lý các trang web đã tải về (loại bỏ các thẻ HTML) để có được dữ liệu là các tài liệu chuẩn. Sau khi có được các tài liệu chuẩn, chúng tôi tiến hành xác định các tài liệu là bản dịch của nhau để tiến hành tách câu và liên kết câu. Bước cuối cùng, chúng tôi xây dựng hệ thống quản lý kho ngữ liệu song song bao gồm các chức năng: thêm mới, chỉnh sửa, xóa các câu song song.

3.1.2. Thiết kế các thành phần hệ thống

3.1.2.1. Xây dựng Module tìm kiếm và xác định các Website đa ngữ

3.1.2.2. Tải các Website đa ngữ

3.1.2.3. Tiền xử lý

3.1.2.4. Xác định các cặp tài liệu là dịch của nhau

3.1.2.5. Tách câu

3.1.2.6. Liên kết câu

3.2. XÂY DỰNG HỆ THỐNG QUẢN LÝ KHO NGỮ LIỆU

SONG SONG

3.2.1. Thực trạng và yêu cầu của hệ thống

3.2.1.1. Phân tích thực trạng

Hệ thống của chúng tôi là một môi trường cộng tác giúp cho cộng đồng có thể phát triển kho ngữ liệu song song Việt – Anh, Việt – Pháp nên nó đòi hỏi phải có đầy đủ các yêu cầu như một môi trường cộng

tác thực sự. Bên cạnh đó hệ thống còn phải đảm bảo tính dễ quản lý và trao đổi giữa các thành viên, tính chia sẻ và dễ sử dụng.

3.2.1.2. Yêu cầu hệ thống

3.2.2. Phân tích thiết kế hệ thống

3.2.2.1. Xác định tác nhân và ca sử dụng

3.2.2.2. Chức năng của hệ thống

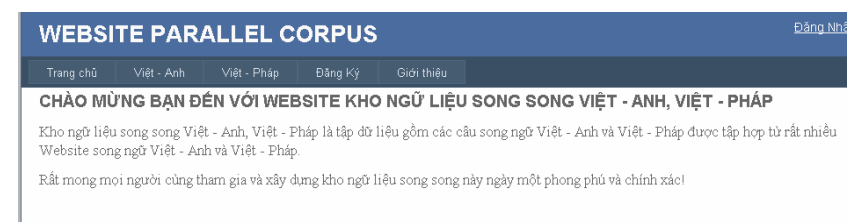
3.2.2.3. Mô tả các ca sử dụng

3.2.2.4. Sơ đồ hoạt động

3.2.3. Thiết kế cơ sở dữ liệu hệ thống

3.3. KẾT QUẢ

3.3.1. Màn hình trang chủ



Hình 3.38. Màn hình trang chủ

3.3.2. Đăng ký thành viên

3.3.3. Đăng nhập

3.3.4. Truy vấn xem dữ liệu

3.3.5. Cập nhật câu song song và đăng bình luận

3.3.6. Chức năng quản trị

3.4. TỔNG KẾT CHƯƠNG 3

Trong chương này chúng tôi đã xây dựng các bước trích lọc dữ liệu từ các Website đa ngữ Việt – Anh, Việt – Pháp. Đặc biệt xây dựng được hệ thống quản lý kho ngữ liệu song song. Qua đó chúng ta thấy khả năng ứng dụng của phương pháp thu thập kho ngữ liệu song song từ các Website đa ngữ là hoàn toàn có thể đưa vào thực tế.

KẾT LUẬN

Với thành tựu và tốc độ phát triển công nghệ như ngày nay thì việc phải vượt qua các rào cản về mặt ngôn ngữ là một nhu cầu thiết và việc ứng dụng các hệ thống dịch tự động trong mọi lĩnh vực là tất yếu. Việc xây dựng một kho ngữ liệu song song từ các Website đa ngữ là điều rất cần thiết để tạo ra CSDL cho hệ thống dịch tự động cũng như việc dạy và học ngoại ngữ cho mọi đối tượng.

1. KẾT QUẢ ĐẠT ĐƯỢC

Về mặt lý thuyết: Tôi đã nghiên cứu các phương pháp thu thập kho ngữ liệu song song từ các Website đa ngữ. Luận văn đã phân tích, đề xuất các bước xây dựng kho ngữ liệu song song từ các Website đa ngữ. Từ đó mở ra một hướng mới trong việc khai thác nguồn tài nguyên đa ngữ khổng lồ từ Internet. Về công nghệ tôi sử dụng .NET và SQL Server 2008 để làm cơ sở phát triển ứng dụng.

Về ứng dụng: Thu thập một số các Website đa ngữ và tải được một số lượng về máy tính. Ngoài ra xây dựng được kho ngữ liệu song song Việt – Anh, Việt – Pháp và hệ thống quản lý cho phép cập nhật,

thêm mới, xóa cũng như bình luận các câu song song; dễ dàng tra cứu hệ thống, tìm kiếm trực tuyến trên Website.

Chúng tôi xây dựng kho ngữ liệu song song này nhằm tạo ra một CSDL mở cho phép mọi người có thể sử dụng trong việc xây dựng hệ thống dịch tự động, trong giảng dạy và học tập ở nhà trường, trong việc nghiên cứu và đánh giá các hệ thống xử lý ngôn ngữ.

2. HẠN CHẾ

Các Module của hệ thống chưa hoàn toàn tự động, một số Module còn cần phải có sự tác động của con người. Giao diện hệ thống Website quản lý kho ngữ liệu song song chưa bắt mắt, chưa đẹp.

3. HƯỚNG PHÁT TRIỂN CỦA ĐỀ TÀI

Từ những nghiên cứu ở trên, chúng tôi cũng mạnh dạn đề xuất các hướng nghiên cứu và phát triển tiếp luận văn trong tương lai như sau:

Thứ nhất, xây dựng cho nhiều cặp ngôn ngữ khác trên thế giới.

Thứ hai, phát triển, nâng cấp giao diện tương tác với người dùng đối với hệ thống quản lý kho ngữ liệu song song để thuận tiện hơn cho người sử dụng.

Thứ ba, xây dựng các Module hoàn toàn tự động, ổn định. Thêm các chức năng mới cho phép nhiều tùy chọn cho các chức năng.

Thứ tư, tích hợp các Module vào một chương trình để dễ sử dụng và quản lý.

-----*