BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

HOÀNG NHƯ QUỲNH

NGHIÊN CỨU XÂY DỰNG KHO DỮ LIỆU SONG NGỮ PHỤC VỤ XỬ LÝ TIẾNG VIỆT

CHUYÊN NGÀNH: **KHOA HỌC MÁY TÍNH**

MÃ SỐ: **60.48.01**

TÓM TẮT 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: **PGS.TS. Võ Trung Hùng**

Phản biện 1: **GS.TS.Nguyễn Thanh Thủy**

Phản biện 2: **PGS.TS.Tăng Tấn Chiến**

Luận văn sẽ được bảo vệ tại Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng ngày 10 và 11 tháng 8 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

Với sự ra đời của máy tính điện tử và nhất là môi trường kết nối Internet toàn cầu đã tạo ra một lượng thông tin khổng lồ đặc biệt đa phần các dữ liệu đều là tiếng Anh. Tuy nhiên lượng thông tin to lớn này vẫn chưa được khai thác hết bởi nhiều lý do và một trong những lý do quan trọng đó là rào cản về ngôn ngữ.

Vấn đề xử lý ngôn ngữ tự nhiên hiện nay rất cần các tài liệu song ngữ, tuy nhiên các tài liệu thường nằm rải rác nhiều nơi dưới nhiều hình thức khác nhau. Do đó tất cả các tài liệu xử lý ngôn ngữ tự nhiên đều dựa vào kho dữ liệu song ngữ ví dụ như dịch tự động, học tiếng Anh, khai thác thông tin trên web,…Vì vậy đòi hỏi một kho dữ liệu song ngữ rất lớn.

Hiện nay trên thế giới có rất nhiều kho dữ liệu song ngữ như Anh – Pháp, Pháp – Anh, Anh – Hoa,… Tuy nhiên, đối với tiếng Việt hiện nay chưa có kho dữ liệu song ngữ nào như vậy được công bố chính thức và chia sẽ cho người sử dụng. Vấn đề đặt ra là làm thế nào để xây dựng được một kho dữ liệu song ngữ Anh – Việt từ các nguồn dữ liệu rải rác.

Để góp phần giải quyết vấn đề trên, chúng tôi đề xuất đề tài: ***"Nghiên cứu xây dựng kho dữ liệu song ngữ phục vụ xử lý tiếng Việt"***.

2. Mục tiêu nghiên cứu

Mục tiêu chính mà đề tài hướng đến là nghiên cứu xây dựng kho dữ liệu chứa các cặp câu Anh – Việt từ các nguồn tài liệu khác nhau như: trang web, từ điển, sách, văn bản,… dưới nhiều định dạng khác nhau, như: XML, TXT, DOC,... và nghiên cứu các nguồn tài liệu như từ điển Lạc Việt, báo tiếng Anh – tiếng Việt, văn bản song ngữ Anh –

Việt,… Để đáp ứng mục tiêu đã nêu, đề tài cần giải quyết những vấn đề chính sau: tìm hiểu về các kho ngữ liệu song song, thu thập các nguồn ngữ liệu song ngữ Anh – Việt, nghiên cứu các giải pháp xây dựng kho dữ liệu song ngữ Anh – Việt để tạo ra được một cơ sở dữ liệu phục vụ cho việc học tiếng Anh, dịch tự động, nghiên cứu xử lý ngôn ngữ tự nhiên,....

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu là các cơ sở dữ liệu của kho dữ liệu song ngữ, các nguồn tài liệu có thể xây dựng nên kho dữ liệu song ngữ.

4. Phương pháp nghiên cứu

Đề tài sử dụng các kỹ thuật tách câu từ một văn bản, bài báo,... Tìm hiểu cách xây dựng kho dữ liệu song ngữ để xây dựng kho dữ liệu song ngữ Anh – Việt.

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

Kho dữ liệu song ngữ Anh – Việt là tài nguyên có giá trị trong việc tạo ra được một cơ sở dữ liệu phục vụ cho việc dạy và học tiếng Anh, dịch tự động, nghiên cứu xử lý ngôn ngữ tự nhiên,...

6. Cấu trúc luận văn

Báo cáo của luận văn được tổ chức thành 3 chương.

Chương 1. Nghiên cứu tổng quan. Trình bày khái niệm về kho ngữ liệu song ngữ, các ứng dụng của kho, nghiên cứu một số kho ngữ liệu song ngữ đang có trên thế giới; nghiên cứu về XML, một số thuật toán về xử lý ngôn ngữ tự nhiên,…

Chương 2. Giải pháp xây dựng kho dữ liệu song ngữ. Chúng tôi trình bày một số giải pháp xây dựng kho ngữ liệu song ngữ.

Chương 3. Phát triển ứng dụng. Trình bày kết quả xây dựng kho dữ liệu từ nhiều nguồn dữ liệu khác nhau.

CHƯƠNG 1: NGHIÊN CỨU TỔNG QUAN

Trong chương này chúng tôi trình bày các vấn đề liên quan đến kho dữ liệu song ngữ, các hệ cơ sở dữ liệu và phương pháp xử lý ngôn ngữ tự nhiên áp dụng khi xây dựng các kho dữ liệu song ngữ.

1.1. Kho dữ liệu song ngữ

1.1.1. Khái niệm

a. Ngữ liệu là những dữ liệu, cứ liệu của ngôn ngữ, tức là những chứng cứ thực tế sử dụng ngôn ngữ. Những chứng cứ sử dụng ngôn ngữ này có thể là của ngôn ngữ nói mà cũng có thể là ngôn ngữ viết. Trong đó ngữ liệu tồn tại dưới dạng ngôn ngữ viết bao gồm nhiều hình thức khác nhau như: dạng giấy, dạng điện tử.

Ngữ liệu chỉ gồm các văn bản của một ngôn ngữ gọi là ngữ liệu đơn ngữ và ngữ liệu của nhiều ngôn ngữ gọi là ngữ liệu đa ngữ.

b. Kho ngữ liệu là một tập hợp các mảnh ngôn ngữ được chọn lựa và sắp xếp theo một số tiêu chí ngôn ngữ học rõ ràng để được sử dụng như một mẫu ngôn ngữ.

Hoặc:

Kho ngữ liệu là một hệ thống tham chiếu dựa trên một bộ sưu tập điện tử của văn bản bao trong một ngôn ngữ nhất định.

c. Kho dữ liệu song ngữ là một kho các cặp văn bản song ngữ được trình bày dưới dạng điện tử, trong đó có mỗi ngôn ngữ là bản dịch của ngôn ngữ kia.

1.1.2. Ứng dụng của kho dữ liệu song ngữ

1.1.2.1. Ứng dụng trong ngôn ngữ học – thống kê

Ngôn ngữ học - thống kê là ứng dụng phương pháp xác suất - thống kê vào việc thống kê, đo, đếm các đối tượng trong ngành ngôn ngữ học.

1.1.2.2. Ứng dụng trong ngôn ngữ học so sánh

Ngôn ngữ học so sánh là so sánh các điểm tương đồng, khác biệt giữa các ngôn ngữ. Để so sánh chúng ta cần có các cứ liệu của các ngôn ngữ mà chúng ta cần so sánh vì vậy việc thu thập, tổng hợp cứ liệu từ các nguồn khác nhau là rất cần thiết.

1.1.2.3. Ứng dụng trong giảng dạy ngoại ngữ

Kho ngữ liệu song ngữ đóng vai trò quan trọng trong việc làm nguồn ngữ liệu và tài liệu sư phạm rất phong phú, làm giàu thêm kiến thức của họ và cũng là công cụ hữu ích trong việc thiết kế giáo trình, sử dụng trong việc dạy và học ngoại ngữ.

1.1.2.4. Ứng dụng trong việc nghiên cứu dịch thuật

Kho ngữ liệu song song có thể giúp phiên dịch để tìm ra sự tương đương giữa ngôn ngữ nguồn và đích. Chúng cung cấp thông tin về tần số của từ, sử dụng cụ thể từ vựng và cú pháp. Giúp phiên dịch để phát triển các chiến lược dịch thuật có hệ thống các từ hay cụm từ hay câu không có tương đương trực tiếp bằng ngôn ngữ đích.

1.1.3. Nghiên cứu một số kho dữ liệu song ngữ trên thế giới

1.1.3.1. British National Corpus (BNC)

Kho ngữ liệu 100.000.000 từ được lấy từ các mẫu văn bản từ nhiều nguồn. Phần ngôn ngữ viết của BNC (90%) được lấy từ các tờ báo, các tạp chí,... Phần ngôn ngữ nói (10%) bao gồm phiên âm chữ viết của các cuộc hội thoại không chính thức và ngôn ngữ nói.

1.1.3.2. Canadian Hansard Corpus (Anh – Pháp)

Kho ngữ liệu với 90 triệu từ Anh – Pháp, là ngữ liệu song song nổi tiếng được trích từ các văn bản của Quốc hội Canada, đã được xuất bản bằng ngôn ngữ chính thức tại Canada là tiếng Anh và tiếng Pháp.

1.1.3.3. JENAAD Japanese-English Parallel Corpus (Anh-Nhật)

Kho ngữ liệu Japanese - English News Article Alignment Data (JENAAD) chứa 150.000 cặp câu. Nguồn gốc của kho ngữ liệu được

lấy từ Yomiuri Shimbun, một trong những tạp chí quốc gia của Nhật Bản, và tờ báo tiếng Anh Daily Yomiuri.

1.1.3.4. PKU 863 (Anh - Trung) của Đại học Bắc Kinh

Kho ngữ liệu song song Anh - Trung PKU trong Dự án 863 của Viện Ngôn ngữ học Tính toán của Trường đại học Peking. Kho ngữ liệu gồm có hơn 200.000 liên kết những cặp câu được lấy từ những văn bản song ngữ có chất lượng (3.066.435 từ tiếng Anh và tiếng Trung Quốc), bao gồm nhiều thể loại và lĩnh vực.

1.2. Một số kỹ thuật sử dụng để xây dựng kho dữ liệu song ngữ

1.2.1. Cơ sở dữ liệu

1.2.1.1. Tổng quan về XML

1.2.1.2. Thuật ngữ

1.2.1.3. Cấu trúc của một file XML

1.2.1.4. Tạo lập một tài liệu XML

1.2.1.5. Những thành phần của một tài liệu XML

1.2.1.6. Kết Luận

1.2.2. Thu thập dữ liệu

Các kho ngữ liệu song ngữ hiện nay thường được chọn lọc từ các nguồn tài liệu như: báo chí, sách, các website song ngữ, ngữ liệu điện tử,... Tuy vậy có một số hạn chế đó là các ngữ liệu song ngữ có sẵn trên mạng Internet đa số đều là các bản dịch thoát ý, hoặc không dịch 1 - 1.

Các nguồn ngữ liệu song ngữ Anh - Việt có thể thu thập:

a. Nguồn từ điển: trong mỗi từ điển, ở mỗi mục từ, thường chứa các ví dụ hướng dẫn sử dụng từ đó, và các ví dụ bằng tiếng Anh này cũng được dịch chính xác (1 - 1) sang tiếng Việt.

b. Ngữ liệu SUSANNE: đây là ngữ liệu điện tử tiếng Anh, gồm khoảng 128.000 từ được rút từ ngữ liệu Brown.

c. Nguồn Internet: đây là nguồn dữ liệu khổng lồ, nguồn ngữ liệu này có lợi thế là chúng đã tồn tại sẵn dưới dạng điện tử, nhưng chỉ có một số ít các trang Web song ngữ là đáp ứng được đúng tiêu chuẩn.

d. Nguồn sách: bao gồm các sách dạy tiếng Anh, các mẫu câu tiếng Anh, sách song ngữ tin học, khoa học kỹ thuật,…

1.2.3. Xử lý ngôn ngữ tự nhiên

Song song với việc thu thập dữ liệu, với các nguồn dữ liệu đầu vào thì cần phải có một số công đoạn xử lý văn bản đầu vào, phân tích, tách đoạn, tách câu,…để đạt được mục đích.

1.2.3.1. Xử lý đầu vào

Các văn bản sẽ được làm sạch, xóa những phần không cần thiết. Các trang web sau khi tải xuống sẽ được trích rút nội dung trang web.

1.2.3.2. Tách đoạn

Tách đoạn nhằm mục đích tách văn bản thành các đoạn và xem văn bản là một khối liên tục các câu.

1.2.3.3. Tách câu

Trong văn bản tiếng Anh, tiếng Việt hay một số ngôn ngữ khác, thông thường người ta dùng dấu chấm (.), chấm than (!), chấm hỏi (?) và một số dấu chấm câu khác để nhận biết kết thúc câu. Tuy nhiên do tính nhập nhằng của dấu báo hiệu kết thúc câu nên việc phân định ranh giới không đơn giản. Ví dụ dấu chấm có thể biểu thị cho một dấu thập phân (1,234.567), một cụm từ viết tắt (Mr., Dr., GS., TS., …), kết thúc câu văn và một số trường hợp như địa chỉ trang web, email…(www.udn.vn hoặc abc@udn.vn). Dấu chấm hỏi hay dấu chấm thang có thể xuất hiện trong dấu ngoặc đơn, ngoặc kép hay ở cuối câu.

1.3. Một số giải thuật trong xử lý ngôn ngữ tự nhiên

1.3.1. Thuật toán liên kết từ

1.3.2. Thuật toán liên kết từ bằng lớp ngữ nghĩa ClassAlign

1.3.3. Thuật toán tách câu

CHƯƠNG 2: GIẢI PHÁP XÂY DỰNG KHO DỮ LIỆU SONG NGỮ

Trong chương này chúng tôi xin trình bày một số giải pháp xây dựng kho dữ liệu song ngữ. Các giải pháp đề xuất bao gồm: Xây dựng kho từ nguồn dữ liệu từ điển, từ nguồn báo điện tử, từ các kho dữ liệu được xây dựng sẵn.

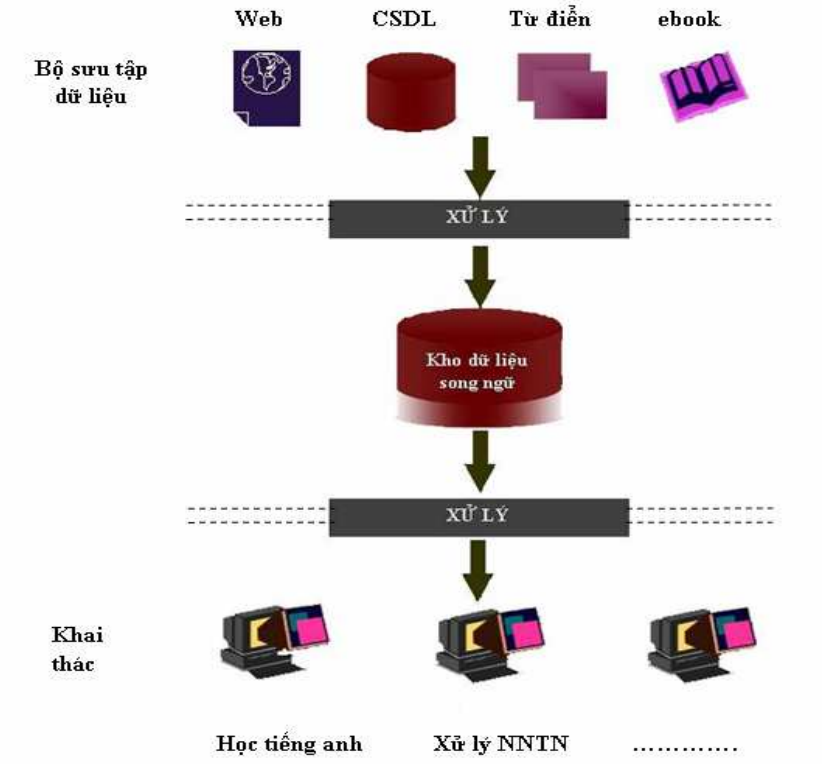
2.1. Giới thiệu

Trong nhiều năm trở lại đây, tầm quan trọng kho ngữ liệu song ngữ được đánh giá rất cao do đó việc xây dựng một kho ngữ liệu song ngữ nhằm đáp ứng nhu cầu về thông tin, về học tập, dịch thuật,… là rất cần thiết. Hiện nay với lượng thông tin trên mạng toàn cầu đa phần là tiếng Anh, tại Việt Nam số lượng kho ngữ liệu song song Anh – Việt không nhiều và không được phổ biến rộng rãi, do đó trong luận văn này chúng tôi đưa ra giải pháp để xây dựng kho ngữ liệu song ngữ Anh – Việt phục vụ xử lý tiếng Việt nhằm đáp ứng nhu cầu sử dụng kho cho giảng dạy, học tập tiếng Anh, dịch máy, xử lý ngôn ngữ tự nhiên,…

2.2. Mô hình tổng thể

Kiến trúc tổng thể của hệ thống bao gồm những thành phần sau:

- Bộ sưu tập dữ liệu: sưu tập các nguồn dữ liệu song ngữ Anh – Việt ban đầu từ ebook, văn bản song ngữ, các trang web song ngữ, từ điển,...
- Tiền xử lý dữ liệu: có thể nhập trực tiếp dữ liệu, xử lý thủ công hoặc hệ thống, chuẩn hóa dữ liệu trước khi đưa vào kho. Việc chuẩn hóa dữ liệu là việc chuyển đổi định dạng dữ liệu thành định dạng tương thích với mục đích của hệ thống.
- Khai thác dữ liệu: những ứng dụng của dữ liệu song ngữ sau khi xử lý.



Hình 2.1. Mô hình tổng thể hệ thống

2.3. Xây dựng kho dữ liệu song ngữ

2.3.1. Các tiêu chí chọn mẫu ngữ liệu

Để bảo đảm được hiệu quả khai thác, đúng mục tiêu nghiên cứu đã đặt ra, chúng ta cần áp dụng 4 tiêu chí trong khi xem xét lấy mẫu ngữ liệu song ngữ Anh-Việt như sau:

a. Chuẩn ngôn ngữ: ngữ liệu tiếng Anh cũng như tiếng Việt đều phải là những câu được xem là chuẩn mực, nghĩa là phải đúng ngữ pháp và được nhiều người chấp nhận hay nhiều người sử dụng.

b. Cách dịch 1 – 1: các ngữ liệu song ngữ Anh-Việt phải thực sự là bản dịch 1 - 1 của nhau, không được dịch thoát ý, dịch tóm lược, dịch tương đương/ đồng nghĩa hay dịch theo kiểu giải thích, diễn giải.

c. Ngữ liệu phải phù hợp với phong cách và lĩnh vực của đối tượng nghiên cứu: Đối tượng nghiên cứu của chúng tôi là các văn bản và các câu thông thường.

d. Ngữ liệu dạng điện tử: ngoài 3 tiêu chuẩn bắt buộc trên, chúng ta sẽ ưu tiên chọn những ngữ liệu song ngữ Anh-Việt nào mà đang tồn tại dưới dạng điện tử.

2.3.2. Chọn nguồn dữ liệu và chuẩn hóa

Trong các nguồn tài liệu thô ta thường thấy các câu ví dụ song ngữ trong các nguồn ngữ liệu khác nhau thì có hình thức trình bày khác nhau. Ví dụ như:

Input-process-output

An input is something put into the system, a process is a series of actions or changes carried out by the system, while an output is something taken from the system.

Đầu vào-xử lý-đầu ra

Một đầu vào được đưa vào hệ thống, một quá trình xử lý gồm hàng loạt hành động hoặc sự sửa đổi được thực hiện bởi hệ thống và cho đầu ra khỏi hệ thống

Hoặc

Dazed, suffering intolerable pain from throat and tongue, with the life half throttled out of him, Buck attempted to face his tormentors.	Đầu choáng váng, họng và lưỡi đau nhức nhối, trong tình trạng đã bị bóp cổ đến gần như ngắc ngoải, Bắc gắng sức đương đầu với những tên hành hạ mình.

Hình 2.2. Ví dụ hình thức trình bày các nguồn dữ liệu khác nhau

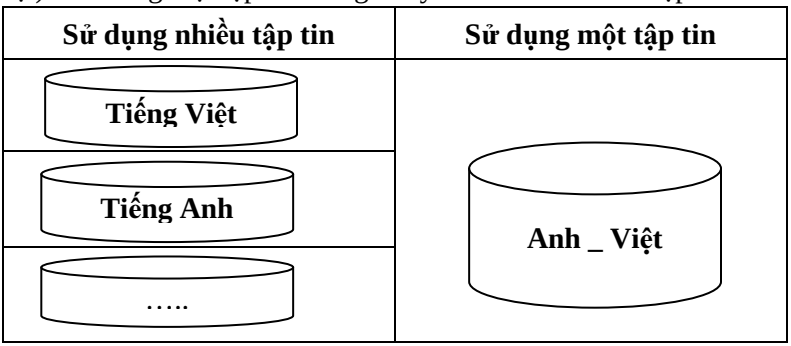
Để chuẩn hoá thành một dạng, một tiêu chuẩn duy nhất. Việc chuẩn hoá ngữ liệu gồm hai nhiệm vụ chính:

1. *Chuẩn hoá dạng ngữ liệu song ngữ Anh - Việt:* đưa về đúng dạng điện tử, định dạng tập tin, mã/font tiếng Việt, chuẩn chính tả.

2. *Liên kết câu (sentence – alignment):* phân ngữ liệu thành từng cặp câu song ngữ Anh - Việt bằng cách đánh dấu xem ứng với mỗi câu tiếng Anh, có câu tiếng Việt nào đi kèm.

2.3.3. Xây dựng cấu trúc kho dữ liệu song ngữ

Về mặt tổ chức lưu trữ dữ liệu chúng tôi chọn việc lưu trữ kho trên XML. Có hai giải pháp để lưu trữ là lưu trữ toàn bộ dữ liệu (Anh, Việt) trên cùng một tập tin đa ngữ hay lưu trữ trên nhiều tập tin:



Hình 2.3. Các giải pháp tổ chức CSDL

Trong trường hợp dữ liệu được lưu trữ trên nhiều tập tin, mỗi ngôn ngữ sẽ được lưu trữ trên một tập tin và có được đánh số chỉ mục giống nhau.

Trong trường hợp này chúng tôi chọn giải pháp lưu trữ trên nhiều tập tin với các lý do sau:

- Khi cần thiết bổ sung ngôn ngữ sẽ dễ dàng, ta chỉ cần tạo ra một tập tin dữ liệu ở ngôn ngữ mà ta muốn cùng cấu trúc sử dụng với các ngôn ngữ khác và lưu trữ song song với các tập tin khác.

- Cấu trúc của tập tin không thay đổi, tất cả các tập tin dữ liệu đều có cùng một cấu trúc và điều này rất có lợi khi lập trình để khai thác các dữ liệu

2.3.4. Các nguồn dữ liệu thu thập

2.3.4.1. Nguồn Từ điển Lạc Việt

Từ điển là một thiết bị, công cụ cho phép lưu trữ thông tin mà qua đó, dựa vào một từ, một cụm từ đơn giản, chúng ta có thể tìm được

nghĩa giải thích, các thông tin liên quan một cách nhanh chóng. Có thể phân chia từ điển thành hai loại lớn:

- Từ điển bách khoa.

- Từ điển ngôn ngữ

Từ điển một ngôn ngữ: Được biên soạn cho một ngôn ngữ cụ thể nào đó ở từng mặt, từng lĩnh vực. Ví dụ: Từ điển giải thích

Từ điển nhiều ngôn ngữ: Được biên soạn trên cơ sở đối chiếu hai hay nhiều ngôn ngữ. Ở đây cũng có thể gồm từ điển đối chiếu phổ thông như: Từ điển Anh – Việt, từ điển toán học Anh – Việt, …

Từ điển điện tử là từ điển được lưu trữ và trình bày trên hệ thống thông tin điện tử. Trong đó có từ điển Lạc Việt là bộ từ điển song ngữ Anh - Việt đầu tiên. Số lượng từ trong phần mềm này rất lớn. Với mỗi từ được tra, chúng sẽ có đầy đủ thông tin về từ loại, ngữ nghĩa, cách phát âm . Tương ứng với mỗi mục từ sẽ có các ví dụ kèm theo khi tra cứu từ trong từ điển Lạc Việt, các mẫu câu ví dụ trong từ điển Lạc Việt là bản dịch 1 - 1 của nhau, vì vậy, các mẫu câu đó là một nguồn dữ liệu chuẩn để xây dựng kho dữ liệu song ngữ Anh–Việt của chúng ta.

Nguồn từ điển Lạc Việt được sử dụng để xây dựng kho dữ liệu song ngữ Anh – Việt được thực hiện qua các bước sau:

- Trích nội dung của các cặp câu ví dụ ứng với mỗi mục từ;

- Tạo cặp kho các cặp câu song ngữ lưu ở tập tin .Doc;

- Xử lý tạo chỉ mục để đưa vào kho.

2.3.4.2. Nguồn Báo điện tử VOV News

Trên mạng Internet có hàng tỷ trang web, một số trong đó là bản dịch của nhau. Web là một nguồn dữ liệu tuyệt vời để xây dựng kho ngữ liệu song song, ít nhất là đối với một số cặp ngôn ngữ. Tuy nhiên, các thủ tục để định vị các văn bản song song trên Web không đơn giản với nhiều lý do sau: 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.

Các website song ngữ thường đặt tên tương tự nhau. Tên trang web luôn gồm có một chuỗi con chung chỉ ra tính song song song của những trang web, cùng đi với một chuỗi con khác được sử dụng như là cờ ngôn ngữ chỉ ra ngôn ngữ của mỗi tài liệu cụ thể. Ví dụ, một trang web tiếng Việt có tên là "vovnews.vn" thì bản dịch tiếng Anh của nó là "english.vovnews.vn".

Để xác định được một trang web là trang web song ngữ thì ở trang ngôn ngữ chính (trang cha) thường có liên kết với các phiên bản ngôn ngữ khác. Trong khuôn khổ luận văn này tôi chọn báo điện tử VOVNews làm nguồn dữ liệu để đưa vào kho dữ liệu song ngữ Anh – Việt cần xây dựng.

VOVNews cũng là một trong những trang web có những bài viết song ngữ Anh - Việt là bản dịch của nhau, tuy nhiên số bài viết là bản dịch của nhau là không nhiều. Và một nhược điểm chung của trang web song ngữ đó là chỉ dịch ý, không phải là bản dịch 1 - 1.

Với nguồn dữ liệu song ngữ này các bước thực hiện bao gồm:

- Tìm kiếm, xác định một cặp trang là bản dịch của nhau;
- Tải các cặp trang web về từ URL;
- Xử lý dữ liệu trích lấy nội dung;
- Tách câu;
- Xử lý để đưa vào kho.

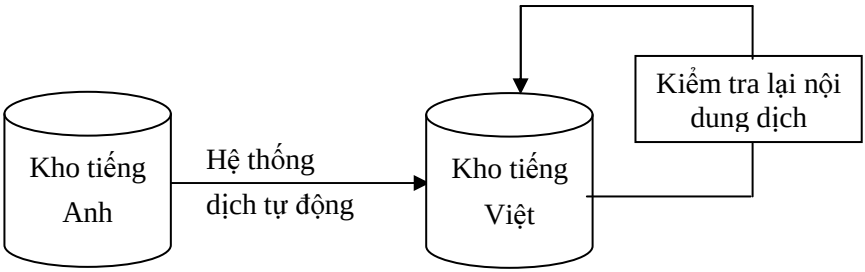
2.3.4.3. Nguồn từ các kho ngữ liệu được xây dựng sẵn

Ngữ liệu huấn luyện là kho ngữ liệu được xây dựng sẵn, các kho ngữ liệu này có thể là đơn ngữ và cũng có thể là song ngữ và từ nhiều

ngôn ngữ khác nhau, các kho ngữ liệu được xây dựng sẵn không không nhiều.

Trên thế giới có rất nhiều kho ngữ liệu song ngữ hoặc đơn ngữ chia sẻ miễn phí cho cộng đồng nghiên cứu. Ví dụ như : kho ngữ liệu song ngữ song song được xây dựng từ sự hỗ trợ của dự án EuroMatrix, ngữ liệu tiếng Anh SUSANNE là ngữ liệu điện tử tiếng Anh được xây dựng bởi một nhóm các nhà ngôn ngữ học - máy tính, kho ngữ liệu Anh – Pháp Canadian Hansard,…

Sử dụng kho ngữ liệu huấn luyện để xây dựng kho ngữ liệu song ngữ Anh - Việt cần phải thông qua một hệ thống dịch tự động để dịch từ ngôn ngữ này sang ngôn ngữ kia.



Hình 2.9. Sơ đồ dịch câu tiếng Anh sang tiếng Việt

Các bước để có được nguồn dữ liệu song ngữ như sau:

- Tìm kiếm các kho ngữ liệu có sẵn;
- Xoá bỏ các tags của XML hoặc dòng trống (nếu có);
- Đưa vào hệ thống dịch tự động;
- Kiểm tra lại nội dung được dịch với sự giúp đỡ của người có chuyên môn;
- Tách câu;
- Xử lý để đưa vào kho.

CHƯƠNG 3: PHÁT TRIỂN ỨNG DỤNG

Trong chương này chúng tôi xin trình bày một số kỹ thuật xử lý nguồn dữ liệu ban đầu thu thập được để xây dựng kho dữ liệu song ngữ bao gồm: kỹ thuật liên kết câu, kỹ thuật cập nhật dữ liệu sử dụng VBA, kỹ thuật trích lọc dữ liệu,… Ở chương này cũng nêu rõ quá trình thực hiện trích từ nguồn từ điển Lạc Việt, từ nguồn báo điện tử VOVNews, từ các kho dữ liệu được xây dựng sẵn qua các kỹ thuật xử lý để xây dựng kho dữ liệu song ngữ.

3.1. Giải pháp xử lý dữ liệu

Trong khuôn khổ luận văn này tôi trình bày một số giải pháp, kỹ thuật xử lý dữ liệu và chuyển đổi từ một số định dạng như rtf, pdf,… sang định dạng XLM .

3.1.1. Kỹ thuật liên kết câu trực tuyến bằng YouAlign

YouAlign là một giải pháp liên kết tài liệu trực tuyến miễn phí, thể truy cập YouAlign ở địa chỉ: http://youalign.com/. Sau khi đăng nhập chúng ta có thể gióng câu giữa hai văn bản song ngữ là bản dịch của nhau. YouAlign cho phép chúng ta download tập tin đã qua xử lý dưới dạng HTML hoặc TMX.

Ưu điểm của YouAlign:

- Cho kết quả gióng câu chính xác với bản dịch của nó.
- Là giải pháp liên kết tài liệu trực tuyến miễn phí
- Hỗ trợ nhiều định dạng.
- Giao diện thân thiện với người dùng.

Nhược điểm:

- Phải sử dụng trực tuyến.
- Tài liệu sau khi download về phải xử lý lại cho phù hợp.

3.1.2. Công cụ cập nhật tài liệu bằng RTF của MS Word

Microsoft word là phần mềm soạn thảo văn bản cao cấp chạy trong môi trường Windows. Word kết hợp nhiều tính năng mạnh như

soạn thảo, định dạng, sử dụng các bộ chương trình tiện ích và phụ trợ giúp tạo các văn bản đặc biệt, macro,... Đặc biệt, để lưu trữ thông tin về cách định dạng sử dụng nhóm định dạng cùng một lúc áp dụng định dạng style.

Ưu điểm

- Ứng dụng ngay tập tin RTF mà không cần phải xây dựng ứng dụng do vậy thời gian triển khai nhanh.

- Việc không xây dựng ứng dụng tra cứu CSDL có nhiều ưu điểm khác như tiết kiệm thời gian tìm hiểu các cấu trúc, các yếu tố liên quan đến việc tổ chức CSDL.

Nhược điểm

- Kích thước tập tin RTF lớn hơn so với các dạng tập tin khác như HTML, XML, DBF khi biểu diễn cùng một lượng thông tin.

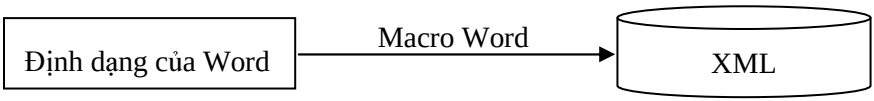
- Có nhiều khó khăn trong việc tìm kiếm.

- Có tính rời rạc vì dữ liệu được lưu trữ trên nhiều tập tin khác nhau và giữa các mục từ không có mối liên hệ về mặt logic.

3.1.3. Kỹ thuật cập nhật sử dụng các macro

Kỹ thuật này được sử dụng cho việc cập nhật kho ngữ liệu song ngữ Anh – Việt. Bằng cách mở Visual Basic Editor trong Word có thể sử dụng VBA viết hay điều chỉnh các macro để định nghĩa các điều khiển ActiveX và tạo ra các ứng dụng trong Word.

VBA là một ngôn ngữ dành cho các macro của Word, các macro ghi nhận sẽ là một thủ tục trong các môđun của VB hay các đề án macro. Một số tiện ích chủ yếu phục vụ sưu tập dữ liệu và chèn thẻ, chỉ mục tương đương cho các câu Anh-Việt, chuyển đổi dạng thức dữ liệu RTF sang XML...



Hình 3.5. Sơ đồ chuyển đổi từ tập tin *. Doc sang tập tin *. XML

Ưu điểm

- Macro tự động tạo ra một dãy các lệnh mà chúng ta đã thao tác, định dạng dữ liệu trong Word và sử dụng các macro bất kỳ lúc nào mà không cần phải khởi động word.

- Sử dụng Macro để xuất các tài liệu Word sang MS Excel, MS Access hay XML.

- Macro tiết kiệm thời gian, công sức và không bị sai sót bằng cách thực hiện một nhóm các lệnh.

- Dữ liệu từ vựng được định dạng theo một cấu trúc nhất định và không mất định dạng nguyên thuỷ như trên các tập tin RTF.

- Việc cập nhật dữ liệu thực hiện một cách dễ dàng, nhanh chóng và có tính mở.

- Dễ dàng viết các câu lệnh VB điều khiển trên cơ sở dữ liệu.

Nhược điểm

- Khi một macro đang thực hiện thì chúng ta không thể can thiệp gì vào cho tới khi macro hoàn tất.

- Nếu thực hiện một macro trong tình trạng sai thì chắc chắn sẽ tốn nhiều thời gian để thực hiện khôi phục trở lại tình trạng ban đầu.

3.1.4. Kỹ thuật trích lọc dữ liệu file html

Thông tin là một tài nguyên cần khai thác và Internet giống như một mỏ tài nguyên khổng lồ. Việc khai thác nội dung của các trang thông tin trên Internet phục vụ cho nhiều mục đích khác nhau, với website song ngữ thì nội dung của cặp trang web song ngữ là bản dịch của nhau là nguồn dữ liệu phong phú để cập nhật kho dữ liệu song ngữ.

Một trang web sau khi được tải về để làm nguồn dữ liệu cập nhật kho, ta cần trích lấy nội dung cần thiết và phải làm sạch, bao gồm:

- Đọc nội dung văn bản đưa về định dạng chuỗi ký tự .

- Hủy bỏ dòng trắng không được hiển thị trên HTML.

- Hủy bỏ các khoảng trắng tab.

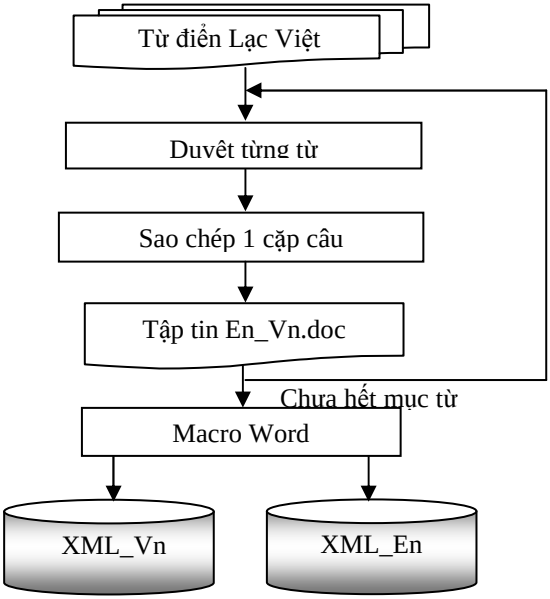
- Hủy bỏ các ký tự trắng liên tiếp trong HTML.
- Hủy bỏ thẻ HEAD.
- Hủy bỏ tất cả JavaScript.
- Thay thế các ký tự đặc biệt như &, <, >, "…
- Kiểm tra và thay thế ngắt dòng (
) hoặc khoản (<p>)
- Loại bỏ tất cả các thẻ HTML.

3.1.5. Một số định dạng đã xử lý

Tập tin Word có đuôi mở rộng .doc và .docx; tập tin Acrobat Reader có đuôi mở rộng .pdf; tập tin html có định dạng html, htm.

3.2. Trích từ từ điển Lạc Việt

Lạc Việt là bộ từ điển song ngữ Anh - Việt phổ biến hiện nay. Số lượng các cặp câu Anh – Việt đi kèm với mỗi từ trong từ điển rất lớn, đồng thời là những cặp câu là bản dịch chuẩn của nhau, là nguồn dữ liệu phong phú để cập nhật kho dữ liệu song ngữ Anh – Việt .



Hình 3.9. Sơ đồ quá trình trích từ Từ điển Lạc Việt

Giải pháp xử lý đưa vào kho dữ liệu song ngữ:

Ở công đoạn này chúng tôi sử dụng đoạn chương trình viết trên VBA để tiến hành chuyển đổi tập tin *.Doc sang định dạng XML và thiết lập các chỉ mục cho các cặp câu Anh – Việt tương ứng.

Sau khi chuyển đổi về dạng XML sẽ nhận được kết quả như sau :

```
<?xml version = "1.0" encoding="UTF-8"
standalone="yes"?>
    <root>
            <sen id=" 1">
               Quả đất thì tròn
            </sen>
            <sen id=" 2">
               Anh ta là giáo viên
            </sen>
    </root>
```

Hình 3.14. Kết quả sau khi chuyển đổi định dạng tập tin và tạo chỉ mục

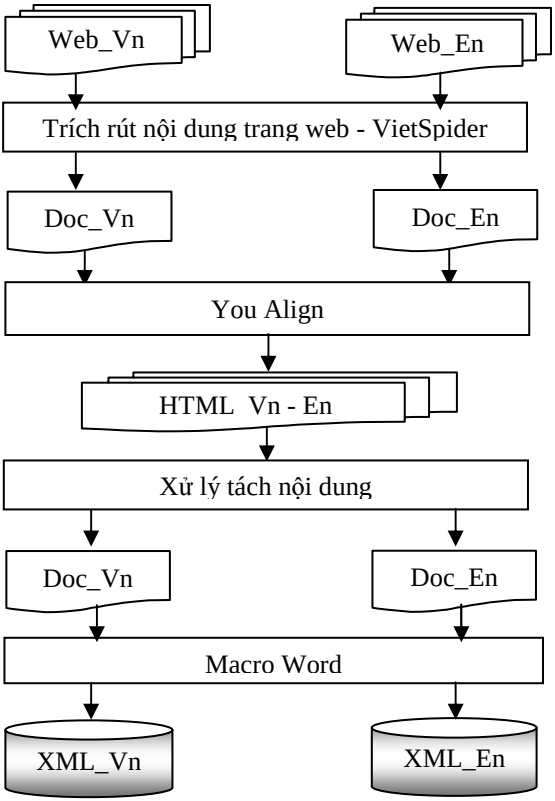
3.3. Trích từ VOV News

Trên World Wide Web tồn tại nhiều dữ liệu, để tìm được hai trang web là bản dịch của nhau tức là nội dung của trang này là bản dịch sang ngôn ngữ khác của nội dung trang kia, ta có thể sử dụng các bộ máy tìm kiếm như Google, Yahoo,… Tuy nhiên khó để xác định được cặp trang web là bản dịch của nhau. Vì vậy, trong khuôn khổ luận văn này tôi chọn một trang web song ngữ Anh – Việt VOV News để sử dụng trong việc xây dựng kho dữ liệu song ngữ.

Do các trang web song ngữ thông thường được tham chiếu lẫn nhau. Để xác định một cặp tin bài Anh – Việt trên trang VOV, ta dựa vào đường dẫn URL của tin bài, tương ứng với mỗi bài viết tiếng Việt

hoặc tiếng Anh, ta sử dụng tiêu đề của bài viết nhờ công cụ dịch của Google để dịch sang ngôn ngữ kia. Tiếp theo tiến hành tìm kiếm nhờ công cụ tìm kiếm trong website của VOV News. Ví dụ "http://vov.vn/Home/Khi-nao-lai-suat-giam/20117/180874.vov" và "http://english.vovnews.vn/Home/When-will-interest-rates-be-lowered-/20117/128494.vov" là bản dịch của nhau, chúng khác nhau ở mục english và nhan đề của bài báo cũng là bản dịch của nhau.

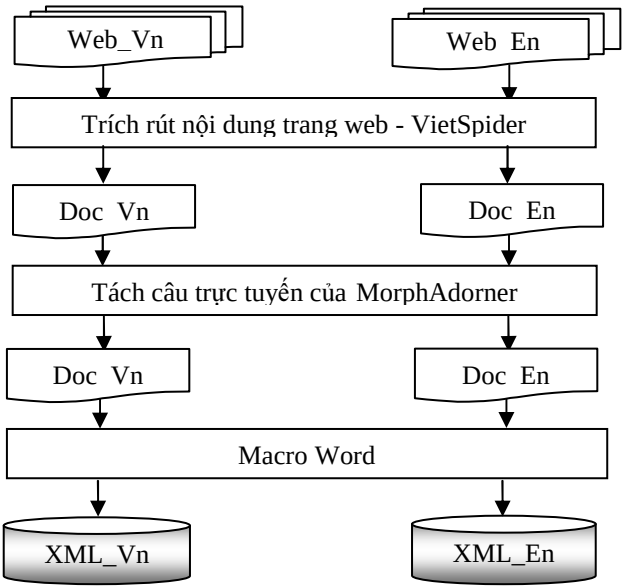
Bước tiếp theo là trích lấy nội dung của trang web. Phần mềm Vietspider là công cụ bóc tách nội dung trang web đúng nghĩa, chúng truy xuất trực tiếp vào nội dung toàn diện rồi tiến hành bóc tách.



Hình 3.16. Sơ đồ quá trình xử lý trích từ trang web sử dụngYouAlign

Nội dung sau khi tách bằng phần mềm Vietspider được đưa vào tập tin .Doc để lưu trữ dữ liệu ban đầu. Tiếp theo ta sử dụng công cụ trực tuyến YouAlign gióng câu giữa hai văn bản song ngữ là bản dịch của nhau sau đó mỗi phần nội dung của mỗi ngôn ngữ được đưa trở lại tập tin .Doc với phần nội dung đã được tách thành các câu riêng biệt. Công đoạn cuối là cập nhật vào kho dữ liệu song ngữ, chúng tôi sử dụng Macro nêu ở phần trên để chuyển đổi và cập nhật dữ liệu.

Trong phần này chúng tôi xin đưa ra một giải pháp tách câu khác đối với những bản dịch 1 – 1 của nhau theo sơ đồ sau:



Hình 3.23. Sơ đồ quá trình xử lý trích từ trang web sử dụng MorphAdorner

Chúng tôi sử dụng công cụ tách câu của MorphAdorner để tiến hành tách thành các câu riêng biệt từ các đoạn trong văn bản. MorphAdorner cung cấp các phương pháp để điều chỉnh văn bản, tách câu,… và có thể sử dụng công cụ tách câu của MorphAdorner trực

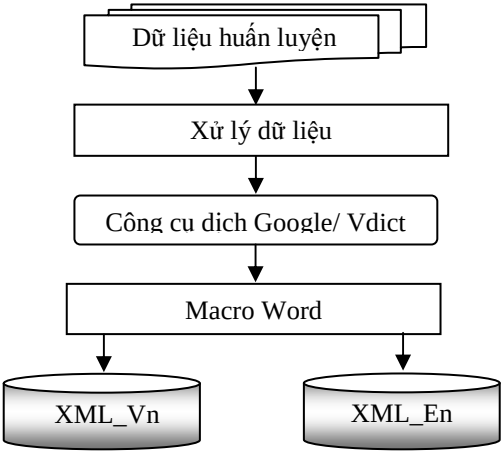
tuyến ở địa chỉ: http://morphadorner.northwestern.edu/morphadorner-/sentencesplitter/example/

Tách câu của MorphAdorner không yêu cầu NSD phải có tài khoản đăng nhập mà cho phép sử dụng trực tiếp. Tuy nhiên việc sử dụng công cụ này vẫn có một số nhược điểm hạn chế đối với tiếng Việt và đòi hỏi NSD phải trực tuyến để sử dụng.

3.4. Trích từ các kho dữ liệu song ngữ Anh - Pháp

Từ một số kho được xây dựng cho phép chia sẻ như: kho ngữ liệu của Nghị viện Châu Âu với 20 ngôn ngữ khác nhau, kho huấn luyện của Hansard,… là những nguồn ngữ liệu đơn ngữ. Một số nguồn dữ liệu được xây dựng sẵn này là các câu tiếng Anh đã được tách, mỗi câu nằm trên một dòng riêng biệt và được lưu trữ dưới định dạng XML. Chúng tôi đã tiến hành xử lý loại bỏ các tags của XML và loại bỏ các dòng trống của những nguồn dữ liệu đơn ngữ được xây dựng sẵn này.

Từ nguồn ngữ liệu này thông qua bộ máy dịch thuật của Google là một công cụ dịch thuật trực tuyến miễn phí được Google cung cấp có thể dịch nhanh văn bản và các trang web,… với nhiều ngôn ngữ. Hoặc sử dụng website dịch tự động trực tuyến Vdict.



Hình 3. 27. Sơ đồ quá trình xử lý nguồn dữ liệu có sẵn

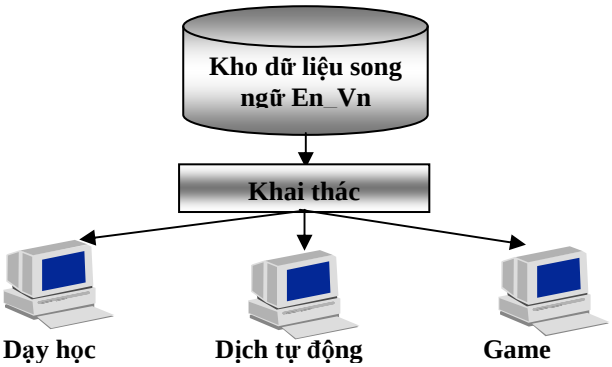
Tất cả những nguồn dữ liệu tiếng Anh và bản dịch tiếng Việt được lưu trữ trong các tập tin .Doc. Tiếp theo chúng tôi sử dụng các công cụ như đã giới thiệu ở phần trên để tiến hành xây dựng, cập nhật kho dữ liệu song ngữ Anh – Việt.

3.5. Một số nguồn dữ liệu khác

Xuất phát từ những hạn chế trong việc tìm kiếm các cặp câu song ngữ Anh-Việt từ các nguồn nói trên. Và để làm phong phú thêm nội dung của kho dữ liệu chúng tôi đã tiến hành tìm kiếm thêm nhiều các cặp câu Anh-Việt từ các nguồn khác như các mẫu truyện, văn bản điện tử được lưu dưới định dạng Pdf và một số website song ngữ khác.

3.6. Khai thác kho dữ liệu song ngữ

Kho ngữ dữ liệu song ngữ chủ yếu được dùng để xây dựng hệ thống dịch tự động, trong nghiên cứu và rất hữu ích trong giáo dục. Kho dữ liệu song ngữ được khai thác trong việc học và giảng dạy ngoại ngữ, trong các trò chơi nhằm trau dồi vốn tiếng Anh. Ngoài ra kho dữ liệu song ngữ còn được sử dụng làm nguồn dữ liệu để sử dụng trong biên soạn phụ đề phim, trong việc xây dựng từ điển, hỗ trợ cho phiên dịch viên,…Với các nhà nghiên cứu, kho dữ liệu song ngữ Anh – Việt được sử dụng trong việc tìm kiếm nghĩa các từ được dịch trong các câu với nhiều bối cảnh khác nhau.



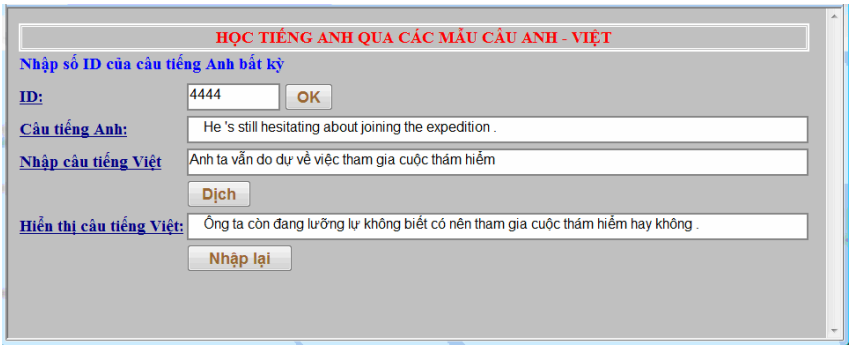
Hình 3.29. Sơ đồ khai thác kho dữ liệu song ngữ Anh – Việt

Để ứng dụng kho dữ liệu song ngữ Anh – Việt vào trong việc dạy và học tiếng Anh, chúng tôi thực hiện một chương trình trò chơi " Học tiếng Anh qua các mẫu câu Anh – Việt" để khai thác kho dữ liệu song ngữ Anh – Việt mà chúng tôi đã xây dựng được. Chương trình cho phép người chơi rèn luyện khả năng dịch qua các mẫu câu Anh – Việt có sẵn.

NSD được yêu cầu nhập mã số ID bất kỳ, mẫu câu tiếng Anh tương ứng sẽ hiển thị ở mục "Câu tiếng Anh".

Chương trình cho phép NSD thể hiện khả năng dịch tiếng Anh của họ bằng cách nhập câu dịch tiếng Việt của người chơi vào mục "Nhập câu tiếng Việt"

Khi NSD muốn so sánh kết quả dịch của mình với bản dịch của chương trình, sau khi nhấn vào nút "Dịch" chương trình sẽ hiển thị câu tiếng Việt tương ứng với câu tiếng Anh mà NSD đang dịch:



Nếu NSD muốn dịch lại hoặc chuyển sang dịch một mẫu câu khác, NSD có thể nhấn nút "Nhập lại" của chương trình "Học tiếng Anh qua các mẫu câu Anh – Việt", chương trình sẽ bắt đầu lại từ đầu.

KẾT LUẬN

Trong quá trình thực hiện luận văn tốt nghiệp này tôi đã thu được nhiều kiến thức về xử lý ngôn ngữ tự nhiên, kho ngữ liệu song ngữ và các vấn đề liên quan đến xử lý dữ liệu. Luận văn trình bày chi tiết các bước cơ bản để thực hiện sao chép ra các tập tin ngữ liệu từ các tập tin định dạng khác nhau ban đầu. Với mục đích có thể khai thác nhiều nguồn dữ liệu khác nhau bằng nhiều công cụ khác nhau. Đồng thời đưa ra các giải pháp, kỹ thuật để xử lý dữ liệu và cập nhật kho dữ liệu song ngữ Anh – Việt.

Tuy nhiên luận văn không tránh khỏi các hạn chế bao gồm: Nguồn dữ liệu song ngữ ở Việt Nam có chất lượng bản dịch không cao, đặc biệt đối với các trang web song ngữ thường dịch ý, tóm lược nội dung của văn bản gốc do đó rất khó khăn cho việc lựa chọn nguồn dữ liệu cũng như canh đoạn, tách câu,…Việc cập nhật kho dữ liệu nói chung vẫn mang tính bán tự động, nhiều công đoạn thủ công. Chưa tìm hiểu kỹ khả năng ứng dụng của kho ngữ liệu vào việc tự động mà chỉ dừng ở mức độ làm dữ liệu phục vụ học tập

Qua quá trình thực hiện luận văn, tôi xin đưa ra một số kiến nghị và hướng phát triển của luận văn như sau: Do nhu cầu nghiên cứu và học tập tiếng Việt của sinh viên nước ngoài, cung như nhu cầu học ngoại ngữ của sinh viên Việt Nam chúng tôi sẽ tiếp tục bổ sung vào nguồn dữ liệu trên, không những chỉ 2 ngôn ngữ Anh – Việt mà có thể thêm nhiều ngôn ngữ khác như Pháp, Trung, Nhật, Hàn,…. Cũng như tìm các giải pháp tối ưu hơn để xây dựng kho dữ liệu hoàn thiện hơn.