BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

LÊ VŨ NGỌC ANH

NGHIÊN CỨU CÁC CÔNG CỤ PHÁT TRIỂN CỦA UNL VÀ KHẢ NĂNG ỨNG DỤNG CHO 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: TS. Nguyễn Trần Quốc Vinh

Phản biện 2: PGS. TS. Lê Mạnh Thạnh

Luận văn được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp Thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 10 Tháng 9 Năm 2011.

Có thể tìm hiểu Luận văn tại:

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

MỞ ĐẦU

1. Lý do chọn đề tài

Những nghiên cứu về dịch tự động đã cho ra đời nhiều công cụ dịch hiệu quả và có thể sử dụng như Google, AltaVista…. nhiều hệ thống đã được đưa vào thương mại hóa như Systran, Reverso, Babylon..... Những công cụ này cho phép tạo ra một "bản dịch nghĩa" - một bản dịch chưa được hoàn chỉnh nhưng giúp chúng ta có thể hiểu được ý nghĩa của văn bản gốc và cần phải chỉnh sửa nhiều để đạt đến một bản dịch hoàn chỉnh. Các hệ thống dịch tự động cho phép dịch rất nhanh và chi phí thấp hơn nhiều so với dịch bằng con người. Tuy nhiên, những hệ thống này đang phải đối mặt với rất nhiều vấn đề như sự đa nghĩa của từ, sự nhập nhằng về ngữ nghĩa, sự phụ thuộc về ngữ cảnh và rất nhiều khó khăn trong sự khác biệt về giải thích các khái niệm.

Có một cách tiếp cận khác tránh rơi vào tình trạng phức tạp của sự đa dạng về ngữ nghĩa; đó là dịch bằng cách sử dụng một ngôn ngữ trung gian (ngôn ngữ biểu đạt riêng cho máy tính). Ngôn ngữ trung gian này cho phép biểu diễn về mặt ngữ nghĩa ở mức đơn giản nhất có thể (giảm thiểu những rắc rối do vấn đề ngữ nghĩa). Một trong những dự án đi theo cách tiếp cận này gọi là Universal Networking Language (UNL). UNL được đề xuất và triển khai thực hiện bởi H.Uchida ở United Nations University, Tokyo, Nhật Bản.

Đối với tiếng Việt, vấn đề đặt ra là làm thế nào để có thể phát triển nhanh nhất hệ thống dịch tự động cho tiếng Việt dựa trên những kết quả sẵn có và UNL là một trong những khả năng để chọn lựa theo hướng này. Vấn đề đặt ra là chúng ta phải nghiên cứu UNL và các bộ công cụ của nó để có thể phát triển nhanh nhất hệ thống dịch tự động cho tiếng Việt dựa trên những kết quả đã có. Được sự gợi ý của PGS. TS. Võ Trung Hùng,

tôi đã chọn đề tài: "*Nghiên cứu các công cụ phát triển của UNL và khả năng ứng dụng cho Tiếng Việt* "

2. Mục đích nghiên cứu

Mục đích là tìm hiểu và trình bày tổng quan về UNL, hệ thống hoạt động và các bộ công cụ của UNL. Trên cơ sở đó, chúng tôi đưa ra khả năng ứng dụng cho tiếng Việt.

3. Đối tượng và phạm vi nghiên cứu

Trong khuôn khổ một luận văn thực nghiệm, chúng tôi chỉ giới hạn nghiên cứu ở việc nắm vững tổng quan ngôn ngữ UNL và các bộ công cụ của nó, giới thiệu tổng quát về các nghiên cứu và giải pháp đã thực hiện để ứng dụng UNL cho tiếng Việt. Trên cơ sở đó, chúng tôi thử nghiệm và đề xuất giải pháp ứng dụng các công cụ phát triển của UNL áp dụng cho tiếng Việt.

4. Phương pháp nghiên cứu

Trong quá trình thực hiện, chúng tôi sử dụng hai phương pháp chính là nghiên cứu tài liệu và thực nghiệm. Với phương pháp đầu tiên, chúng tôi tiến hành thu thập và nghiên cứu các tài liệu có liên quan đến đề tài. Phương pháp tiếp theo là nghiên cứu các công cụ UNL sẵn có, tiến hành thử nghiệm trên các công cụ UNL sẵn có và đề xuất giải pháp ứng dụng cho tiếng Việt. Cuối cùng là đánh giá kết quả và nêu hướng phát triển của đề tài.

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

Báo cáo của đề tài đã trình bày tổng quan về UNL, giới thiệu các công cụ và hệ thống hỗ trợ UNL, sau đó thử nghiệm, đánh giá và đề xuất giải pháp ứng dụng các công cụ phát triển của UNL cho tiếng Việt. Kết quả này sẽ tạo tiền đề cho việc nhanh chóng xây dựng thành công hệ thống dịch tự động đa ngữ cho tiếng Việt trong tương lai.

6. Cấu trúc của luận văn

Báo cáo luận văn được tổ chức thành ba chương. Chương đầu chúng tôi giới thiệu phần nghiên cứu tổng quan về UNL và các bộ công cụ của nó. Chương hai là giới thiệu trình bày tổng quan về về các nghiên cứu và giải pháp đã thực hiện để ứng dụng UNL cho tiếng Việt. Chương ba là tiến hành thử nghiệm trên một công cụ hỗ trợ UNL, đánh giá và đề xuất một số ứng dụng UNL cho Tiếng Việt, triển vọng của đề tài. Cuối cùng là kết luận và nêu hướng phát triển của đề tài.

CHƯƠNG 1
TỔNG QUAN VỀ UNL VÀ CÁC NGHIÊN CỨU ĐỂ ÁP DỤNG UNL CHO TIẾNG VIỆT

Trong chương này, chúng tôi trình bày tổng quan về ngôn ngữ của UNL, hệ thống UNL và giới thiệu một số công cụ phát triển của UNL

1.1. Tổng quan về ngôn ngữ UNL

1.1.1. Khái niệm

UNL là từ viết tắt của "Universal Networking Language". Nó là ngôn ngữ máy tính cho phép máy tính có thể truy cập thông tin và tri thức mà không bị rào cản ngôn ngữ. Nó là một ngôn ngữ giả có khả năng mô phỏng thế giới ngôn ngữ tự nhiên của con người trong giao tiếp. Kết quả là nó cho phép mọi người có thể biểu diễn tất cả các tri thức từ ngôn ngữ tự nhiên. Nó cũng cho phép máy tính giao tiếp, vì thế cung cấp cho mọi người các cấu trúc ngôn ngữ để phân bố, nhận và hiểu thông tin đa ngôn ngữ.

UNL biểu diễn thông tin hoặc tri thức dưới dạng mạng ngữ nghĩa với cấu trúc đa đồ thị. Khác với ngôn ngữ tự nhiên, sự biểu diễn của UNL là

không nhập nhằng. Trong mạng đa ngữ nghĩa của UNL, các nút biểu diễn các khái niệm và các cạnh biểu diễn mối quan hệ giữa các khái niệm.

Từ khi UNL là ngôn ngữ của máy tính, nó có tất cả các thành phần của ngôn ngữ tự nhiên. UNL bao gồm UW - Từ vựng, Relation - Quan hệ, Attributes - Thuộc tính, and UNL Knowledge Base - Kiến thức cơ bản. Nó tạo ra các từ biểu diễn các khái niệm gọi là "Universal Word" gọi tắt là UW, UW chứa các từ vựng của UNL. Nó liên kết nội với các từ vựng khác tạo thành câu. Những liên kết này gọi là "relation" - mối quan hệ, nó chỉ định vai trò của mỗi từ trong câu. Những ngụ ý của người nói có thể được diễn tả thông qua "Attribute" - Thuộc tính.

"UNLKB" cung cấp những định nghĩa ngữ nghĩa của từ vựng. UNLKB định nghĩa mỗi quan hệ có thể có giữa các khái niệm bao gồm các quan hệ phân cấp và các kỹ thuật tham chiếu dựa trên các quan hệ bao gồm lẫn nhau giữa các khái niệm. Vì thế UNLKB cung cấp nền tảng ngữ nghĩa của UNL để chắc chắn nghĩa của biểu thức UNL là không nhập nhằng.

1.1.2. Biểu thức UNL

1.1.3. Các quan hệ

1.1.4. Từ vựng UNL

1.1.5. Phân loại từ vựng UNL

1.1.6. Thuộc tính UNL

1.1.7. Biểu thức UNL

1.2. Tổng quan về hệ thống UNL

1.2.1. Quá trình EnConvertor

1.2.2. Quá trình DeConvertor

1.2.3. Dictionary - Từ điển

1.3. Các nghiên cứu để áp dụng cho UNL-tiếng Việt

1.3.1. Giải pháp dịch tiếng Việt thông qua hệ thống trung gian hỗ trợ UNL

Giới thiệu

Dựa vào việc tìm hiểu một cách có hệ thống về khái niệm UNL, hệ thống hoạt động của UNL. Nghiên cứu đã đề xuất ứng dụng UNL cho tiếng Việt thông qua hai mô hình như sau:

Mô hình 1: Chuyển đổi thủ công văn bản Tiếng Việt sang ngôn ngữ UNL và từ đó sử dụng các công cụ hỗ trợ dịch từ UNL sang các ngôn ngữ khác như Tiếng Anh, Tiếng Nga, Tiếng Tây Ban Nha, Tiếng Ý.

Ví dụ: từ một câu tiếng Việt: “Tôi có thể giúp gì cho ông không ?”, ta sẽ chuyển nó sang dạng UNL:

agt(help(icl>do).@polity.@interrogative.@entry, I)

obj(help(icl>do).@entry.@polity.@interrogative, you)

Và từ đây ta có thể dịch nó sang những ngôn ngữ (hiện nay là 15 ngôn ngữ) đã được hỗ trợ bởi UNL như tiếng Anh, tiếng Pháp, tiếng Nhật,…

Phương pháp thực hiện

Cách 1: Xây dựng kho dữ liệu các câu Tiếng Việt - Tiếng Anh - UNL (ứng dụng hệ thống ETAP3 để chuyển từ tiếng Anh sang UNL). Sau đó sử dụng các trang web dịch trực tuyến để dịch các câu UNL sang ngôn ngữ cần. (Ví dụ Tiếng Nga, Nhật).

Cách 2: Chuyển đổi thủ công văn bản Tiếng Việt sang ngôn ngữ UNL.

Mô hình 2: Phát triển các công cụ hỗ trợ như :

+ Xây dựng công cụ hỗ trợ quá trình Mã hóa - EnConverter :

- Xây dựng từ điển các từ, các luật văn phạm, từ điển định nghĩa các khái niệm cơ bản của Tiếng Việt.

- Xây dựng các luật mã hóa, các luật phân tích từ trong câu.

- Khi chuỗi đầu vào được nạp thì EnConverter sẽ tiến hành phân tích các từ trong câu, mỗi từ được xem như là một nút, nạp luật mã hóa và tiến hành kiểm tra luật. Áp dụng luật mã hóa cho danh sách các nút. Quá trình xử lý của ứng dụng luật là để tìm ra luật thích hợp và áp dụng trên danh sách nút để tạo chức năng cú pháp và mạng UNL sử dụng các nút trong cửa sổ phân tích. Nếu một chuỗi xuất hiện trong cửa sổ, hệ thống sẽ xây dựng từ điển từ và áp dụng luật lên các phần tử từ. Trong trường hợp, nếu một từ đáp ứng đủ các điều kiện yêu cầu cho cửa sổ của luật, từ này sẽ được lựa chọn và ứng dụng luật tiếp tục. Quá trình xử lý này sẽ tiếp tục cho đến khi chức năng cú pháp và mạng UNL được hoàn thành và chỉ còn lại các phần tử nút trong danh sách nút.

- Cuối cùng EnConverter hiển thị mạng UNL thành file dữ liệu đầu ra là quan hệ nhị phân theo định dạng của biểu thức UNL.

+ Xây dựng công cụ hỗ trợ quá trình Giải mã - DeConverter : Gồm 3 thành phần

- Thành phần đầu tiên dùng để chuyển biểu thức UNL thành đồ thị.

- Thành phần thứ hai chuyển đồ thị thành một số cây

- Thành phần thứ ba dùng phương pháp đệ quy duyệt từ trên xuống qua các đỉnh để dịch mỗi cây con và kết quả là một câu hoàn chỉnh.

Nhận xét

Đối với mô hình 1, theo cách 1 thì ưu điểm là nhanh chóng ứng dụng UNL mà không cần phải phát triển bất cứ công cụ bổ sung nào. Nó phù hợp với việc phổ biến nhanh những dữ liệu cơ bản và thiết yếu (hướng dẫn du lịch, các mẫu hội thoại đơn giản, quảng cáo…) ra nhiều thứ tiếng đã hỗ trợ bởi UNL. Hạn chế là phải có đội ngũ am hiểu ngôn ngữ UNL để chuyển những dữ liệu đang có sang UNL. Đối với cách 2, do UNL được xây dựng dựa trên từ điển các từ của tiếng Anh nên chỉ có một số ít các từ Tiếng Việt có thể định nghĩa thành từ Tiếng Anh để máy chủ có thể hiểu từ đấy và chọn từ Tiếng Nga thích hợp. Do đó, để có thể thực hiện được công cụ có thể mã hóa từ Tiếng Việt sang các ngôn ngữ khác thì ta cần xây dựng bổ sung các định nghĩa của các khái niệm tương ứng giữa Tiếng Việt – Tiếng Anh.

Đối với mô hình 2: Đối với mô hình ứng dụng 2, ưu điểm là tạo ra một hệ thống dịch tự động đa ngữ hoàn chỉnh cho tiếng Việt; đó là phát triển các mô-đun dịch tiếng Việt - UNL và UNL - tiếng Việt. Tuy nhiên, với mô hình này thì cần phải bỏ ra nhiều công sức để nghiên cứu, phát triển dữ liệu từ điển, ngữ pháp và các mô-đun dịch trên cơ sở nền tảng đã có của UNL

1.3.2. Giải pháp xây dựng từ điển UNL-tiếng Việt

Giới thiệu

Để ứng dụng nhanh chóng hệ thống UNL phục vụ dịch đa ngữ cho tiếng Việt; nhiệm vụ quan trọng nhất là tích hợp được tiếng Việt vào UNL.

Để làm được việc này, chúng ta cần phát triển mô-đun dịch xuôi (tiếng Việt - UNL) và dịch ngược (UNL - tiếng Việt). Mỗi mô-đun bao gồm nhiều công đoạn nhỏ khác nhau, trong đó một phần quan trọng phục vụ cho dịch từ động để đưa ra những bản dịch chính xác vẫn là cơ sở dữ liệu từ điển.

Giải pháp này đưa ra dựa trên việc nghiên cứu cấu trúc từ điển Anh - Việt theo định dạng Dict. Hiện nay, www.dict.org đã xây dựng một định dạng từ điển rất dễ sử dụng, định dạng này đã được một số cá nhân sử dụng để xây dựng những bộ từ điển khá lớn. Có nhiều bộ từ điển thông dụng đã được cộng đồng phát triển. Nghiên cứu này sử dụng bộ từ điển Anh - Việt của tác giả Hồ Ngọc Đức (http://www.informatik.uni-leipzig.de/~duc/Dict/) để trích phần nội dung tiếng Việt. Về chuẩn chính tả tiếng Việt vẫn tuân theo chuẩn chính tả như trong từ điển Hoàng Phê. Về mã tiếng Việt, tác giả sử dụng bộ mã Unicode. Bên cạnh đó, nghiên cứu cũng đã sử dụng từ điển UNL - FR (hơn 39.000 từ) do nhóm GETA (Groupe d'Etudes pour la Traduction Automatique) xây dựng.

Phương pháp thực hiện

Qua nghiên cứu cầu trúc từ điển UNL-FR và từ điển Anh-Việt theo chuẩn Dict của tác giả Hồ Ngọc Đức, nghiên cứu đã đề xuất các bước xây dựng từ điển UNL - tiếng Việt như sau:

- Lấy một mục từ tiếng Pháp trong từ điển UNL-FR

- Lấy headword và các thuộc tính từ loại đi cùng như CATV, CATN, CATADJ…của mục từ tiếng Pháp đó.

- Lấy một mục từ trong từ điển Anh – Việt

- Lấy headword mục từ đó và các thuộc tính đi cùng với như động từ, danh từ, tính từ,…

- So sánh 2 headword vừa lấy từ 2 từ điển, nếu giống nhau thì tùy theo từ loại là danh từ, động từ, tính từ,…thì gán nghĩa tiếng Việt vào nội dung mục từ tiếng Pháp tương ứng → được 1 mục từ UNL - tiếng Việt → lưu mục từ vừa tạo vào cơ sở dữ liệu từ điển UNL - tiếng Việt

- Quá trình sẽ lặp lại liên tục cho đến khi khai thác hết các mục từ trong từ điển UNL-FR.

Nhận xét

Việc xây dựng từ điển UNL - tiếng Việt bằng phương pháp so sánh các headword dựa vào từ điển UNL-FR và Anh - Việt đã tạo được một số lượng khá lớn từ vựng (247.763 từ). Những headword trong từ điển UNL-FR không tìm thấy trong từ điển Anh - Việt (bảng 1) là 36.85% có thể giải thích bởi các nguyên nhân như sau:

- Hệ thống chưa xử lý hết cấu trúc chi tiết bên trong của mỗi mục từ trong từ điển Anh - Việt. Ví dụ trong từ điển UNL-FR có headword là “hurry_up”, nhưng trong từ điển Anh - Việt headword chỉ có “@hurry”, còn “hurry_up” là các chi tiết bên trong của động từ “hurry”.

- Từ trong Anh - Việt chưa đầy đủ hoặc chưa khai thác hết các thuộc tính nằm trong các CAT của UNL-FR.

1.3.3. Giải pháp xây dựng môi trường cộng tác để phát triển từ điển UNL-tiếng Việt

Giới thiệu

Giải pháp đưa ra là xây dựng môi trường cộng tác trao đổi và chia sẽ kiến thức để phát triển từ điển UNL - tiếng Việt, góp phần vào việc đưa ứng dụng UNL vào xử lý tiếng Việt. Nhiệm vụ chính là nghiên cứu về dịch dự động: các khái niệm, lịch sử phát triển, các phương pháp, những hạn chế và một số ứng dụng. Nghiên cứu về sử dụng ngôn ngữ trục (Pivot Language) trong xử lý ngôn ngữ tự nhiên, các vấn đề liên quan đến từ điển, tìm hiểu chi tiết về UNL và từ điển trong UNL và nghiên cứu về môi trường hợp tác trên mạng.

Phương pháp thực hiện

Bước 1: Thiết kế kho dữ liệu

Kho dữ liệu được thiết kế dựa vào file thành lập từ nhiều nguồn chỗ chứa dữ liệu đã được sắp xếp theo dạng điện tử của phù hợp với cấu trúc mà tổ chức UNL thế giới sử dụng. Kho dữ liệu được thiết kế để thuận tiện cho việc báo cáo và phân tích cũng như trích xuất để sử dụng góp phần làm nền tảng cho việc phát triển các công cụ dịch tự động về sau. Kho dữ liệu được thiết kế mục đích ở đây là tập trung vào việc lưu giữ dữ liệu. Những dữ liệu này sẽ được kiểm tra và đưa vào dữ liệu từ điển để có thể sử dụng làm nền tảng phát triển cho các hệ thống deconvertor cho Tiếng Việt.

Nghiên cứu này đã đề xuất giải pháp xây dựng kho dữ liệu dựa vào từ điển Anh - Việt theo định dạng Dict của tác giả Hồ Ngọc Đức (http://www.informatik.uni-leipzig.de/~duc/Dict/) để trích phần nội dung tiếng Việt. Về chuẩn chính tả tiếng Việt vẫn tuân theo chuẩn chính tả như trong từ điển Hoàng Phê và sử dụng bộ mã Unicode. Kết hợp với từ điển UNL-FR (hơn 39.000 từ) do nhóm GETA xây dựng.

Bước 2: Xây dựng môi trường cộng tác

Xây dựng một website là một môi trường cộng tác để phát triển từ điển UNL-Tiếng Việt có đầy đủ các yêu cầu như một môi trường cộng tác thực. 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.

Nhận xét

Hệ thống xây dựng từ điển trên mạng cộng tác giúp nhiều người có thể chung sức để nhanh chóng xây dựng nên một cơ sở dữ liệu từ điển UNL-Tiếng Việt có giá trị. Nó là một từ điển mở nên mang tính chất dân chủ giúp nhiều tác giả thuộc nhiều lĩnh vực chuyên môn khác nhau có thể cộng tác với nhau và đưa ra nhiều bình luận hữu ích hỗ trợ trong việc lựa chọn cập nhật nội dung của từ điển sao cho chính xác.

Tuy nhiên, việc xây dựng từ điển UNL-Tiếng Việt trên mạng cộng tác cũng gặp một số vấn đề hạn chế như việc kiểm soát bài viết là rất khó, cũng như tính chuyên môn trong số cộng tác viên sẽ không đồng đều, sự khách quan sẽ tùy thuộc rất nhiều vào người quản lý chính, và công tác quản lý xét duyệt trong môi trường cộng tác mạng thì rất khó khăn.

CHƯƠNG 2
CÁC CÔNG CỤ VÀ HỆ THỐNG HỖ TRỢ UNL

Trong chương này, chúng tôi sẽ trình bày một số công cụ và hệ thống hỗ trợ UNL. Ở mỗi công cụ chúng tôi sẽ trình bày tổng quát và nhận xét khả năng áp dụng cho tiếng Việt.

2.1. Hệ thống ETAP- 3

2.1.1. Giới thiệu

ETAP-3 là môi trường NLP đa tiện ích mà nó được được hình thành vào năm 1980 và là sản phẩm của Institute for Information Transmission Problems, Russian Academy of Sciences (Apresjan et al. 1992a, b, Boguslavsky 1995). ETAP-3 được trên lý thuyết Ngữ nghĩa - Văn bản (Meaning - Text) của Igor' Mel'čuk và the Integral Theory of Language của Jurij Apresjan. ETAP-3 là phần mềm chủ yếu để phục vụ cho môi trường nghiên cứu đa ngữ hơn là phần mềm có tính thương mại. Trọng tâm chính của việc nghiên cứu với ETAP-3 là mô hình tính toán của ngôn ngữ tự nhiên. Tất cả các ứng dụng của NLP trong ETAP-3 phần lớn dựa trên ba giá trị logic và sử dụng ngôn ngữ chuẩn cho miêu tả đa ngữ, FORET.

ETAP-3 có tổ chức các kiến thức ngôn ngữ học. Nghĩa là dữ liệu ngôn ngữ (văn phạm và từ điển) được dựa trên khái niệm từ phần mềm sử dụng để xử lý chúng. Theo đó, kiến thức ngôn ngữ không bị phân tán trong mã phần mềm và vì thế dễ hiểu, dễ sử dụng và sửa chữa.

2.1.2. Các chức năng của ETAP-3

Các module chính NLP của ETAP-3 như sau :

- ✓ Hệ thống dịch máy (Machine Translation System)
- ✓ Giao diện ngôn ngữ tự nhiên để truy vấn dữ liệu
- ✓ Hệ thống diễn giải các câu tương đương.
- ✓ Công cụ sửa lỗi cú pháp
- ✓ Công cụ hỗ trợ máy tính học ngôn ngữ.
- ✓ UNL Deconverter và Enconverter

Những tính năng quan trọng nhất của môi trường ETAP-3 và trong các module như sau:

- ✓ Phương pháp dựa trên luật (Rule-Based Approach)
- ✓ Phương pháp phân tầng (Stratificational Approach)
- ✓ Phương pháp kế thừa (Transfer Approach)
- ✓ Sự độc lập cú pháp (Syntactic Dependencies)
- ✓ Phương pháp từ vựng (Lexicalistic Approach)
- ✓ Hệ thống dịch phức tạp (Multiple Translation)
- ✓ Nguồn tài nguyên của ngôn ngữ có thể mở rộng tối đa.

2.1.3. ETAP-3 và UNL

ETAP-3 là hệ thống NLP dựa trên nguồn tri thức ngôn ngữ dồi dào, nó có thể được dễ dàng mở rộng và ứng dụng cho các ứng dụng khác. Phương pháp của hệ thống ETAP-3 nhằm xây dựng cầu nối giữa UNL và một trong những cách biểu diễn nội của ETAP, tên là NormSS (Normalized Syntactic Structure), và theo cách này sẽ liên kết UNL với các ngôn ngữ khác dưới dạng biểu diễn văn bản.

Mức biểu diễn NormSS là thích hợp nhất cho việc thiết lập phù hợp với UNL, với biểu thức UNL. Tầm quan trọng của chúng như sau :

- ✓ Cả biểu thức UNL và NormSS giữ vị trí trung gian giữa giao diện và việc biểu diễn ở mức ngữ nghĩa. Chúng phù hợp cả ở mức cú pháp. Ở mức này, nghĩa của các phần tử từ vựng không được phân tích thành gốc và mối quan hệ giữa các phân tử từ vựng là độc lập với ngôn ngữ.

- ✓ Các nút của cả biểu thức UNL và NormSS đều là các phần tử nhỏ nhất và không có cấu trúc cú pháp
- ✓ Các nút chứa các đặc điểm riêng (gọi là thuộc tính)
- ✓ Các cạnh của cả hai cấu trúc là cấu trúc không đối xứng phụ thuộc.

2.2. Công cụ CWL Conversion Framework

2.2.1. Giới thiệu

CWL Conversion Framework là một công cụ cung cấp sự mã hóa qua lại giữa các loại định dạng CWL.unl, CWL.cdl và CWL.rdf. Nó là một ứng dụng web độc lập được viết bằng ngôn ngữ Java/JSP, DHTML/Ajax và VML.

2.2.2. Các chức năng chính

- ✓ Phân tích dữ liệu đầu vào
- ✓ Xây dựng đối tượng đồ thị với các nút để chỉ các phần tử và các cung chỉ các quan hệ.
- ✓ Phát sinh ra các hình thức xem khác nhau (UNL, CDL, RDF, Graphical)

Các chức năng cơ bản này được cài đặt như thư viện của Java mà nó có thể được sử dụng cho các ứng dụng khác.

2.2.3. Các kiểu hiển thị

Các đối tượng đồ thị được trả về có thể hiển thị ở các dạng khác nhau nhưng nhìn chung chúng phản ảnh mối quan hệ tương tự như nhau từ dữ liệu đầu vào.

+ Graph View

+ UNL View

+ CDL (Concept Description Language) View

+ RDF View

2.3. Hệ thống Unl Explorer

2.3.1. Giới thiệu

UNL Explorer là một ứng dụng cho phép người sử dụng hoặc các nhà phát triển xem hoặc phát triển cơ sở dữ liệu UNL (UNL Database). UNL Database lưu trữ thông tin của UNL trong đó thông tin chính là các từ vựng UWs (Universal Words). Các từ vựng (Uws) được lưu trữ trong từ điển UNL và mỗi từ vựng được miêu tả bằng biểu thức UNL. Dựa trên UNL Database, UNL Explorer cho phép người sử dụng tìm kiếm thông tin sử dụng từ vựng UWs hoặc từ một ngôn ngữ tự nhiên nào đó. Nó sẽ hiển thị các kết quả trong UNL hoặc một ngôn ngữ tự nhiên mong muốn bằng cách truy cập vào hệ thống UNL. Hệ thống giải mã (Deconverter) của UNL sẽ giải biểu thức thông tin UNL ra một ngôn ngữ tự nhiên mong muốn. Nó cũng cung cấp một số chức năng cho các nhà phát triển thêm hoặc sửa đổi thông tin trong UNL Database trong ngôn ngữ mẹ đẻ của họ. Trong trường hợp này, công cụ UNL Editor là rất cần thiết để tạo ra các biểu thức UNL từ các ngôn ngữ tự nhiên. Kiến trúc của UNL Database cho phép phát triển của nó sẽ được thực hiện bởi nhiều nhà phát triển từ các ngôn ngữ và nền văn hóa khác nhau.

2.3.2. Cấu trúc của UNL Database

UNL Database gồm có 2 phần: UNLKB cung cấp những định nghĩa ngữ nghĩa của từ vựng và UNL Document chứa nội dung thông tin các tài liệu UNL.

2.3.3. Cấu trúc của UNL Explorer

UNL Explorer có hai kiểu là UNL Explorer Editor và UNL Explorer Viewer.

2.3.4. Cài đặt

Tất cả các tập tin và thư mục phải được lưu trữ trong cùng một thư mục với cái tên "C:\UNLExplorer". Nếu sử dụng ở một ổ đĩa khác, thì tên ổ đĩa "C" phải được thay thế bởi tên ổ đĩa đó. Trong trường hợp này, phải thay thế lại tất cả tên ổ đĩa trong tập tin UNLExpV.ini và UNLExpV.ini.

2.3.5. Chức năng của UNL Explorer

Tập tin chương trình chính của UNL Explorer là UNLExpE.exe. UNLExpV.exe là một chương trình chỉ dành cho người xem.

2.4. Công cụ Word Dictionary Builder

2.4.1. Giới thiệu

Word Dictionary Builder là một công cụ để tạo nên chỉ mục của từ điển từ dữ liệu văn bản. Chỉ mục từ điển có thể được sử dụng ở cả 2 quá trình mã hóa và giải mã.

2.4.2. Cách sử dụng và định dạng từ điển từ văn bản

2.5. Công cụ UNL PLATFORM

2.5.1. Giới thiệu

UNL Platform là một UNL dựa trên tài liệu đa ngôn ngữ phát triển ứng dụng web. Nó cung cấp cho người dùng một môi trường tích hợp để người dùng có thể xây dựng tài liệu UNL (UNL Documents) từ ngôn ngữ tự nhiên UNL và ngược lại. UNL Platform tích hợp tất cả các công cụ cần thiết của hệ thống UNL và cung cấp các chức năng khác nhau để giúp người sử dụng trong xây dựng UNL và tài liệu ngôn ngữ đích. Tùy thuộc vào nhu cầu và mục đích của người sử dụng, UNL Platform cung cấp nhiều cấp độ khác nhau của các chức năng để đáp ứng nhu cầu.

2.5.2. Đặc điểm

Hiện UNL Platform chỉ mới hỗ trợ cho tiếng Anh và tiếng Nhật, trong tương lai sẽ là tiếng Trung và một số tiếng khác.

2.6. Công cụ JIBIKI

2.6.1. Giới thiệu

Jibiki là một môi trường chung cho các văn bản trực tuyến và truy vấn tất cả các loại từ điển: thuật ngữ, từ điển song ngữ, từ vựng đa ngôn ngữ cơ sở dữ liệu,…Nó đã được phát triển bởi Mathieu Mangeot (Université de Savoie, Pháp) và Gilles Sérasset (Université de Grenoble 1, Pháp), hiện nay có thêm sự tham gia của Francis Brunet - Manquat, nhóm GETA của phòng thí nghiệm CLIPS ở Grenoble, Pháp.

Được xây dựng bằng công nghệ Java và những công cụ mã nguồn mở độc quyền. Nó dựa trên Enhydra, một máy chủ web động và Postgres, cơ sở dữ liệu quan hệ. Giao diện hiện nay là bằng tiếng Anh, tiếng Estonia, Pháp, Đức và Nhật Bản. Người dùng cũng có thể dễ dàng thêm một ngôn

ngữ mới. Một số thuận lợi cho việc giao tiếp giữa các cộng đồng người sử dụng là diễn đàn, danh sách phân phối.

2.6.2. So sánh với các công cụ khác

2.6.3. Một số dự án sử dụng Jibiki

- Papillon Project
- GDEF Project
- LexALP Project

2.6.4. Một số chức năng

- Tra cứu từ điển
- Quản lý các nhiệm vụ

2.7. Công cụ UW GATE

2.7.1. Giới thiệu

Công cụ UW Gate cung cấp cho người dùng phương tiện để truy cập vào UNL Ontology và từ điển UW thông qua Internet. Sử dụng công cụ UW Gate, người dùng có thể tìm kiếm những từ mong muốn, mối quan hệ các từ, từ tương đương của ngôn ngữ tự nhiên… Người dùng cũng có thể định nghĩa hoặc đăng ký từ mới tương đương với ngôn ngữ tự nhiên. Từ mới được đưa vào vị trí thích hợp trên hệ thống UW Gate bằng cách làm theo hướng dẫn của UW Gate, để chúng có thể làm cho các chức năng trong bản thể UNL thực hiện tốt hơn.

2.7.2. Chức năng

- Ngôn ngữ hỗ trợ: Hiện tại UW Gate hỗ trợ 20 ngôn ngữ thông dụng trên thế giới như Anh, Pháp, Đức, Nhật, Tây Ban Nha, Ấn Độ, Hàn Quốc, Ý, Nga… trong đó có tiếng Việt.
- Chức năng đăng nhập để sử dụng
- Chức năng tìm từ và sửa đổi
- Chức năng tìm từ nâng cao

2.7.3. Nhận xét

Công cụ UW Gate cho phép các nhà phát triển để kiểm tra, chỉnh sửa, thêm, hoặc bỏ qua các mục từ UW thông qua Internet. Công việc xây dựng từ điển là một quá trình lâu dài và đòi hỏi cần có sự tham gia của một cộng đồng qua tâm đến lĩnh vực này, vì vậy đây một công cụ rất hữu ích trong quá trình xây dựng từ điển UNL - tiếng Việt.

2.8. Công cụ Universal Parser

2.8.1.Giới thiệu

Universal Parser tạo ra các biểu thức UNL từ câu đầu vào không cần sử dụng thông tin ngữ pháp của ngôn ngữ phụ thuộc, mà chỉ sử dụng các chú thích của ngôn ngữ độc lập. Câu được đưa vào Universal Parser phải được chú thích với chú thích UNL. Universal Parser phân tích các chú thích của câu đầu vào bằng cách sử dụng quy tắc Universal Parser và một từ điển UW.

2.8.2. Chức năng

Để sử dụng đúng Universal Parser, bao gồm cả hình thức của từ trong từ điển UW, hoặc thay đổi tất cả các hình thức chuyển từ câu đầu vào

thành các dạng cơ bản nếu từ điển UW chỉ chứa các hình thức cơ sở là cần thiết. Thay vào đó, chỉ đơn giản bằng cách mở rộng quy tắc Parser Universal bao gồm một tập hợp các quy tắc phân tích hình thái học của một ngôn ngữ, một chú thích dựa trên hình thái tùy chỉnh Parser của một ngôn ngữ có thể được dễ dàng thực hiện.

Có thể xem thêm thông tin tại http://www.undl.org/unlsys/uparser/UP.htm. và có thể sử dụng tại www.undl.org/up/ (thời điểm hiện tại UP đang được nâng cấp nên không sử dụng được)

2.9. Kết luận

Cho đến nay, đối với tiếng Anh và một số ngôn ngữ phổ biến khác trên thế giới thì việc xử lý tự động ngôn ngữ tự nhiên bằng hệ thống UNL đã đạt được những thành tựu đáng kể. Hiện đã có rất nhiều công cụ được phát triển để hỗ trợ cho việc nghiên cứu và ứng dụng dịch máy bởi hệ thống UNL. Điều quan trọng là làm thế nào nhanh chóng áp dụng hệ thống UNL phục vụ dịch đa ngữ cho tiếng Việt bằng cách nghiên cứu và ứng dụng những công cụ đã có.

CHƯƠNG 3
THỬ NGHIỆM CÁC CÔNG CỤ CỦA UNL

Qua việc trình bày tổng quan về UNL, các nghiên cứu đã thực hiện về UNL áp dụng cho tiếng Việt ở chương 1; cũng như các công cụ hỗ trợ UNL ở chương 2. Trong chương này, chúng tôi sẽ tiến hành thử nghiệm và đánh giá 3 trong số các công cụ đã giới thiệu ở chương 2 là công cụ Jibiki, UNL Explorer và UNL Platform.

3.1. Công cụ JIBIKI

3.1.1. Giới thiệu

3.1.2. Thử nghiệm

3.1.3. Các chức năng chính

3.2. Công cụ UNL EXPLORER

3.1.1. Giới thiệu

3.1.2. Thử nghiệm

3.1.3. Các chức năng chính

3.3. Công cụ UNL PLATFORM

3.1.1. Giới thiệu

3.1.2. Thử nghiệm

3.1.3. Các chức năng chính

3.4. Nhận xét và hướng nghiên cứu

Qua việc trình bày các chức năng và thử nghiệm 3 công cụ là công cụ Jibiki, công cụ UNL Explorer, công cụ UNL Platform chúng tôi xin đưa ra các nhận xét và giải pháp ứng dụng như sau:

3.4.1. Ưu điểm

So với giải pháp xây dựng môi trường cộng tác để phát triển từ điển qua mạng Internet đã trình bày ở chương 2, chúng ta nhận thấy rằng ứng dụng công cụ Jibiki để xây dựng từ điển là lựa chọn thích hợp nhất. Vì không phải tồn nhiều thời gian và công sức xây dựng công cụ cộng tác xây dựng từ điển trên mạng Internet. Công cụ này đã được nhiều dự án về ngôn ngữ sử dụng và được kiểm chứng là tốt. Trong thử nghiệm của mình, chúng tôi dựa trên trang web dự án từ điển Papillon có địa chỉ tại

http://aximag.fr/pivax/Home.po. Công cụ này tương đối dễ sử dụng và có thể truy cập bằng các phương pháp như web, di động và từ điển dạng Dictd.

Đối với công cụ UNL Explorer với phiên bản 2010 đã được tích hợp nhiều công cụ khác của UNL như UNL Editor, UW Gate... đồng thời hỗ trợ hơn 20 ngôn ngữ thông dụng trong đó có cả tiếng Việt nên đây là một thuận lợi rất lớn để ứng dụng công cụ này vào việc nghiên cứu UNL cho tiếng Việt .

Công cụ UNL Platform là một công cụ tuyệt vời để dùng trong hệ thống Mã hóa và giải mã UNL. Tuy nhiên hiện nay chỉ mới tích hợp cho các ngôn ngữ như tiếng Anh, Nhật, Trung.

3.4.2. Nhược điểm

Do UNL hoàn toàn mới mẻ đối với tiếng Việt nên tài nguyên và các nghiên cứu còn rất hạn chế. Các nghiên cứu cũng như các công cụ hỗ trợ cho tiếng Việt không nhiều. Mới chỉ dừng lại ở việc xây dựng các từ điển tiếng Việt – UNL. Ngoài ra, do các máy chủ chỉ hoạt động khi đang thử nghiệm hoặc trong quá trình thực hiện dự án nên việc thử nghiệm thỉnh thoảng gặp trở ngại vì không truy cập được máy chủ hệ thống. Đồng thời muốn trở thành thành viên sử dụng các công cụ và hệ thống UNL bắt buộc thành viên đó phải có nhiều đóng góp cho cộng đồng UNL.

3.4.3. Hướng nghiên cứu

Qua 3 công cụ chúng tôi đã trình bày ở trên, ta thấy có thể hoàn toàn nghiên cứu và ứng dụng các công cụ của UNL cho tiếng Việt. Tuy nhiên, hiện nay do việc nghiên cứu về UNL cho tiếng Việt vẫn còn ít nên tài

nguyên, công cụ chưa nhiều. Để có thể áp dụng nhanh chóng các công cụ này thì chúng ta có thể kế thừa những kết quả đạt được của các tổ chức nghiên cứu UNL cho tiếng Pháp. Bên cạnh đó, chúng ta phải nhanh chóng tham gia vào cộng đồng UNL để cùng nghiên cứu và chia sẻ những thành quả có được.

Đối với việc xây dựng từ điển UNL cho tiếng Việt thì do đặc thù của tiếng Việt nên cần xử lý tính nhập nhằng và tăng độ chính xác của kho dữ liệu. Khắc phục những hạn chế trên để có được một hệ thống hoàn chỉnh cần phải tiếp tục nghiên cứu thêm về ngôn ngữ Việt Nam và cấu trúc tiếng Việt để đảm bảo sự đúng đắn cho dữ liệu từ điển được xây dựng. Bên cạnh đó cần xây dựng thêm môi trường cộng tác tốt hơn có hỗ trợ chat trực tuyến, hỗ trợ xử lý dữ liệu thô sang cấu trúc UNL bên cạnh đó cần phải tiếp tục sửa giao diện người dùng sao cho dễ dàng trực quan hơn đối với người sử dụng.

Để nhanh chóng xây dựng hệ thống dịch đa ngữ cho tiếng Việt cần phải có một hướng đi đúng đắn và kế thừa những kết quả tốt nhất đã có. Như đã nói ở trên, để ứng dụng được hệ thống UNL chúng ta cần xây dựng nhiều mô-đun khác nhau, đây là các công việc tốn nhiều thời gian và tiền bạc. Những nghiên cứu và thử nghiệm trên các công cụ sẵn có của UNL mà chúng tôi đưa sẽ là một hướng tiếp cận khác để nhanh chóng xây dựng một hệ thống dịch tự động cho tiếng Việt.

KẾT LUẬN

Việc nghiên cứu các công cụ của UNL và ứng dụng cho tiếng Việt mặc dù vẫn còn một số hạn chế nhưng đã đạt được một số thành công nhất định. Kết quả lớn nhất mà chúng tôi đạt được qua đề tài là đã nghiên cứu trình bày một cách có hệ thống về UNL, một số công cụ hỗ trợ của UNL và giới thiệu các nghiên cứu đã thực hiện để ứng dụng cho tiếng Việt. Trong quá trình thử nghiệm, chúng tôi đã tiến hành thử nghiệm và đánh giá 3 công cụ của UNL là công cụ Jibiki, UNL Explorer và UNL Platform. Với những kết quả đạt sẽ cho chúng ta có một cái nhìn tổng quan hơn để có giải pháp tiếp cận nhanh nhất các công cụ sẵn có của UNL trong việc xây dựng hệ thống dịch tự động cho tiếng Việt.

Việc nghiên cứu dịch tự động ứng dụng hệ thống UNL cho tiếng Việt chưa được phổ biến ở trong cũng như ngoài nước, nên hầu hết các trang web cũng như các máy chủ ngôn ngữ chưa hỗ trợ cho tiếng Việt. Chúng tôi may mắn kế thừa những kết quả tốt nhất có được từ tổ chức nghiên cứu dịch tự động tiếng Việt ở Pháp. Và chúng tôi chỉ tập trung giới thiệu các bộ công cụ hỗ trợ từ điển sẵn có của UNL và tiến hành thử nghiệm trên đó.

Để nhanh chóng xây dựng hệ thống dịch đa ngữ cho tiếng Việt cần phải có một hướng đi đúng đắn và kế thừa những kết quả tốt nhất đã có. Những nghiên cứu và thử nghiệm mà chúng tôi đưa sẽ là một hướng tiếp cận khác để nhanh chóng xây dựng một hệ thống dịch tự động cho tiếng Việt.