Kỷ yếu Hội nghị Khoa học công nghệ Quốc gia lần thứ XVI về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Đà Nẵng, ngày 28-29/9/2023
DOI: 10.15625/vap.2023.0090

XÂY DỰNG KHO NGỮ LIỆU ĐA NGỮ NHỜ VÀO UNL

Võ Trung Hùng[1], Phan Thị Lệ Thuyền[2], Ninh Khánh Chi[3]

[1]Trường Đại học Sư phạm Kỹ thuật - Đại học Đà Nẵng
[2]Trường Đại học FPT - Campus Quy Nhơn
[3]Trường Đại học Công nghệ thông tin và Truyền thông Việt Hàn - Đại học Đà Nẵng

vthung@ute.udn.vn, thuyenptl@fe.edu.vn, nkchi@vku.udn.vn

TÓM TẮT: *Bài báo này trình bày việc xây dựng kho dữ liệu đa ngữ nhờ vào một ngôn ngữ gọi là UNL (Universal Networking Language). UNL được định nghĩa gồm có các thành phần tương ứng như là một ngôn ngữ tự nhiên và nó có thể biểu diễn mọi thông tin của ngôn ngữ ở dạng có cấu trúc mà không gây nhập nhằng về ngữ nghĩa. Với đặc tính này, ngôn ngữ UNL có thể sử dụng như là một ngôn ngữ trung gian (ngôn ngữ trục) để chuyển đổi qua lại giữa các ngôn ngữ tự nhiên thay vì dịch trực tiếp giữa hai ngôn ngữ mà chúng ta thường hay sử dụng. Việc sử dụng ngôn ngữ UNL trong hệ thống dịch đa ngữ có hai ưu điểm lớn: UNL không có sự nhập nhằng về ngữ nghĩa và giảm số lượng cặp dịch từ n*(n-1)/2 xuống 2*n. Chúng tôi thử nghiệm với 106.434 câu, kết quả đầu ra cho thấy dịch qua UNL tốt hơn so với dịch trực tiếp bằng Google Translator (đánh giá dựa trên hai phương pháp NIST và BLEU).*

Từ khóa: *UNL, Universal networking language, Automatic translation, Multilingual corpus.*

I. GIỚI THIỆU

Internet đã trở thành một phần không thể thiếu để kết nối, chia sẻ và tìm kiếm tài liệu trong cuộc sống hiện đại. Tuy nhiên với lượng thông tin khổng lồ được tạo ra mỗi ngày bởi con người, hầu như chúng ta không thể khai thác hết bởi nhiều lý do và một trong những lý do quan trọng nhất là rào cản về ngôn ngữ. Hiện có hai giải pháp chính để giải quyết vấn đề này: Thứ nhất là phát triển các hệ thống, các ứng dụng, các nguồn dữ liệu đa ngữ để người sử dụng có thể lựa chọn ngôn ngữ mà họ muốn khi sử dụng; Thứ hai là ứng dụng các phần mềm dịch tự động để dịch trực tiếp các giao diện, các nội dung từ ngôn ngữ hiện có sang ngôn ngữ mà người sử dụng chọn lựa.

Dịch tự động bắt đầu từ thế kỷ 17 và hiện nay nhiều hệ thống dịch đa ngữ được xây dựng với sự hỗ trợ của Deep Learning đã nâng độ chính xác của kết quả khiến cho không ít chuyên gia phải đặt ra nghi vấn về khả năng trong tương lai không xa. Tuy nhiên để có kết quả như vậy thì chúng ta cũng cần có nguồn dữ liệu khá lớn. Hiện nay trên thế giới hiện đang sử dụng hơn 5.000 ngôn ngữ có chữ viết, việc phát triển một hệ thống dịch đa ngữ cho từng cặp ngôn ngữ là vô cùng khó khăn và nhất là những ngôn ngữ có số lượng người dùng ít. Một trong những giải pháp là dịch qua ngôn ngữ trung gian, hướng tiếp cận này giảm chi phí xây dựng phần mềm dịch từ (n*(n-1)) xuống còn (2*n) và đây là một giải pháp tốt đối với các cặp ngôn ngữ thiếu tài nguyên hoặc không tương đồng cấu trúc ngữ pháp [1].

Giải pháp sử dụng một ngôn ngữ tự nhiên (như tiếng Anh, tiếng Tây Ban Nha, tiếng Pháp,…) làm ngôn ngữ trung gian được triển khai trong các hệ thống dịch đa ngữ. Vì vậy có nhiều kho ngữ liệu song ngữ lớn (như tiếng Anh, tiếng Tây Ban Nha, tiếng Pháp,…) được xây dựng, đó là một lợi thế cho các hệ thống dịch đa ngữ mới. Tuy nhiên, với phương pháp dịch hai lần thông qua ngôn ngữ thứ ba chất lượng bản dịch không cao vì không khử được tính nhập nhằng của từ loại trong ngôn ngữ tự nhiên. Đến nay, hướng tiếp cận này thường sử dụng cho các cặp ngôn ngữ không tương đồng về cấu trúc ngữ pháp hoặc khan hiếm nguồn tài nguyên dữ liệu nhưng tính chính xác không cao.

Thay vì sử dụng ngôn ngữ tự nhiên làm ngôn ngữ trung gian, UNL là một giải pháp lựa chọn. UNL là ngôn ngữ nhân tạo, nó có các thành phần như là một ngôn ngữ tự nhiên. Hiện nay UNL cũng được mã hóa sang các ngôn ngữ khác như Tây Ban Nha, tiếng Nga, tiếng Nhật,…và ngược lại [6]. Hệ thống dịch tự động đa ngữ bao gồm nhiều máy chủ ngôn ngữ khác nhau được dịch thông qua ngôn ngữ trục là UNL. Mỗi máy chủ ngôn ngữ sẽ đảm nhận hai chức năng, đó là dịch một văn bản từ ngôn ngữ địa phương sang ngôn ngữ UNL gọi là quá trình mã hóa và dịch ngược lại gọi là quá trình giải mã.

Trong bài báo này, chúng tôi trình bày kết quả nghiên cứu và thực nghiệm để xây dựng một kho dữ liệu đa ngữ dựa trên nền tảng của hệ thống UNL. Phương pháp chúng tôi đề xuất là sử dụng UNL như ngôn ngữ trục để khi có dữ liệu từ một ngôn ngữ nào đó thì chỉ cần chuyển nó sang ngôn ngữ UNL và khi cần dữ liệu này ở ngôn ngữ nào thì chỉ cần dịch từ UNL sang ngôn ngữ đó. Kết quả chúng tôi đạt được trong quá trình nghiên cứu gồm: một kho dữ liệu gồm 106.434 câu tương ứng trong 3 ngôn ngữ Anh, Pháp và Việt Nam; xây dựng bộ từ điển UNL - tiếng Việt, phát triển các công cụ EnCovie (dịch một câu tiếng Việt sang ngôn ngữ UNL) và DeCovie (dịch một câu từ ngôn ngữ UNL sang tiếng Việt).

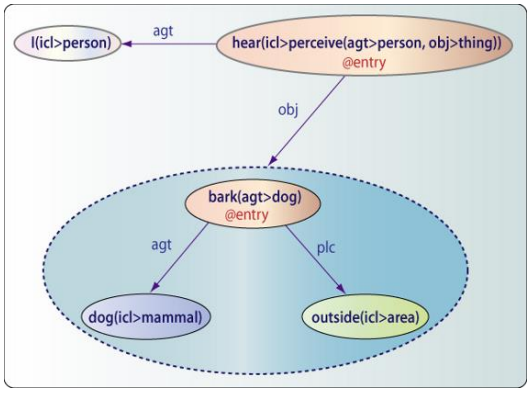
II. CÁC NGHIÊN CỨU LIÊN QUAN

A. UNL

Năm 1996, UNL được phát triển bởi một viện nghiên cứu của United Nations University, Tokyo, Nhật Bản [2] nhằm mã hóa, lưu trữ thông tin độc lập với ngôn ngữ tự nhiên mà nó biểu diễn. Khác với ngôn ngữ tự nhiên, sự biểu diễn thông tin của UNL là không nhập nhằng về ngữ nghĩa. Thông tin UNL biểu diễn dưới dạng mạng ngữ nghĩa với

cấu trúc đa đồ thị, các nút biểu diễn các khái niệm còn các cạnh biểu diễn các mối quan hệ giữa các khái niệm. Các khái niệm được định nghĩa trong UNL gọi là từ vựng (UW), các từ vựng được liên kết với với nhau để tạo thành biểu thức UNL. Các liên kết này được gọi là quan hệ (Relation) nhằm xác định vai trò của mỗi từ vựng trong biểu thức. Ý nghĩa chủ quan của người nói trong câu nguồn sẽ được thể hiện qua thuộc tính (Attributes) trong biểu thức UNL. Ngoài ra một thành phần được dùng để định nghĩa ngữ nghĩa của từ vựng gọi là kiến thức cơ sở UNL (UNLKB), UNLKB đảm bảo chắc chắn nghĩa của từ vựng không nhập nhằng.

Ví dụ câu tiếng Anh *"I can hear a dog barking outside"* sẽ được biểu diễn dưới dạng Graph của UNL như sau:



Hình 1. Biểu diễn dạng Graph của UNL

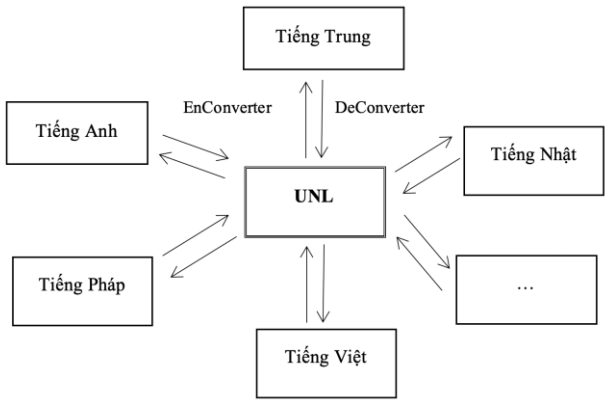
Quan hệ nhị phân của của biểu thức UNL được viết như sau:

```
{unl}
   agt(hear(icl>perceive(agt>person,obj>thing)).@ability.@entry, I(icl>person).@topic)
   obj(hear(icl>perceive(agt>person,obj>thing)).@ability.@entry, :01)
   agt:01(bark(agt>dog).@progress.@entry, dog(icl>mammal).@indef)
   plc:01(bark(agt>dog).@progress.@entry, outside(icl>area))
{/unl}
```

Trong biểu thức UNL ở trên, `agt`, `obj` và `plc` là các **quan hệ**. `I(icl>person)`, `hear(icl>perceive(agt>person, obj>thing))`, `dog(icl>mammal)`, `bark(agt>dog)` và `outside(icl>area)` là các **UW**. `@ability`, `@entry`, `@indef`, `@progress` và `@topic` là các **thuộc tính**.

B. Hệ thống UNL

Một cách ngắn gọn, hệ thống UNL bao gồm nhiều máy chủ ngôn ngữ khác nhau. Mỗi máy chủ ngôn ngữ được cài đặt riêng cho từng ngôn ngữ và đăng ký kết nối với máy chủ UNL để thực hiện việc gửi yêu cầu dịch hoặc nhận lại kết quả [3]. Mỗi máy chủ ngôn ngữ sẽ đảm nhận 2 chức năng đó là chuyển văn bản được viết trong ngôn ngữ tự nhiên sang văn bản được viết trong ngôn ngữ UNL gọi là **EnConverter** và dịch ngược lại được gọi là **DeConverter**.



Hình 2. Hệ thống dịch tự động đa ngữ UNL

C. Kho dữ liệu đa ngữ

Hiện nay, con người đang sử dụng nhiều ngôn ngữ khác nhau và chính sự khác biệt này tạo ra rào cản rất lớn cho việc tiếp cận thông tin. Chính vì sự đa dạng về ngôn ngữ và quá trình toàn cầu hóa đang diễn ra mạnh mẽ nên vấn đề cấp thiết đặt ra hiện nay là làm thế nào để những người nói hoặc viết bằng những ngôn ngữ khác nhau có thể hiểu được nhau dễ dàng hơn. Đối với các ngôn ngữ lớn (ngôn ngữ có nhiều người sử dụng và/hoặc được sử dụng bởi những quốc gia có tiềm lực mạnh về kinh tế, khoa học, kỹ thuật) đã có nhiều kho dữ liệu chất lượng được xây dựng. Chúng ta có thể dễ dàng tìm thấy các kho dữ liệu bằng tiếng Anh, tiếng Pháp, tiếng Hoa, tiếng Nhật… Đặc biệt, có rất nhiều các kho dữ liệu song ngữ Anh - Pháp, Anh - Hoa, Anh - Nhật,… Ngược lại, đối với những ngôn ngữ như tiếng Việt thì những nghiên cứu về nó chưa nhiều, rời rạc và đặc biệt là sự thiếu vắng các kho dữ liệu lớn về khối lượng và đảm bảo

về chất lượng để phục vụ công tác nghiên cứu và phát triển các ứng dụng. Vì vậy, việc nghiên cứu, xây dựng một kho dữ liệu đa ngữ phục vụ cho xử lý tiếng Việt là một vấn đề cần thiết, cấp bách đặt ra hiện nay.

Các tác giả [4] đã tiến hành xây dựng kho dữ liệu đa ngữ (Anh, Pháp, Việt) phục vụ xử lý ngôn ngữ tự nhiên gồm 106.434 câu được trích từ các nguồn:

Bảng 1. Thống kê kho dữ liệu gốc

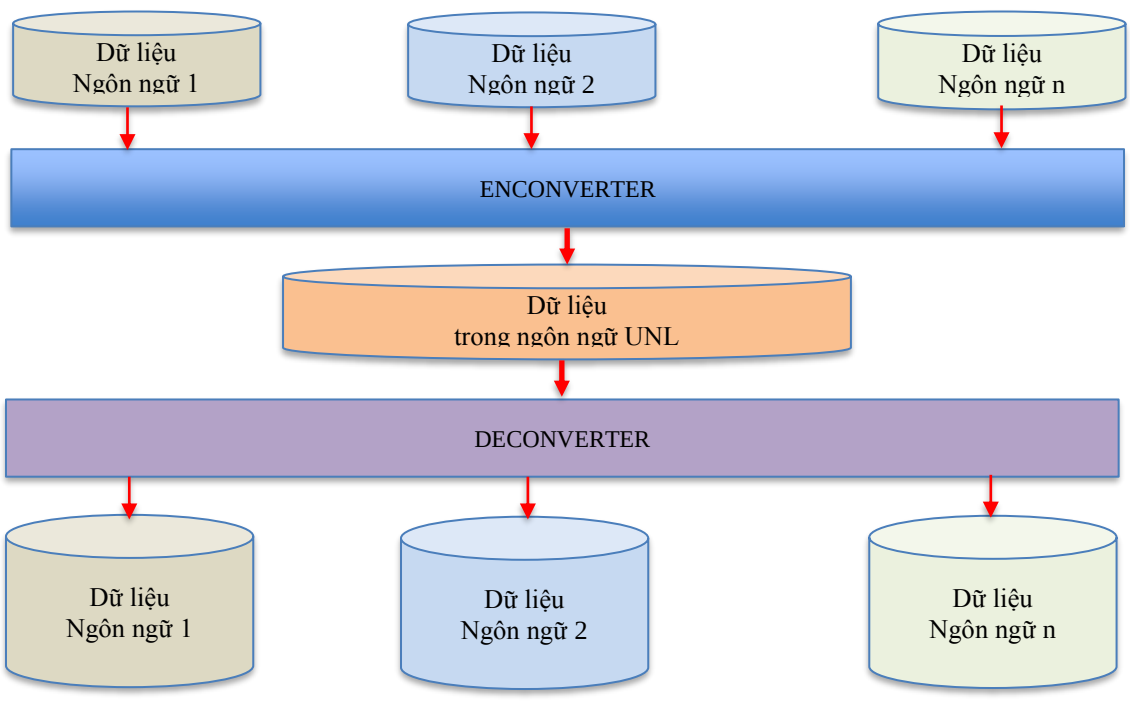
STT	Nguồn gốc	Ngôn ngữ gốc	Số lượng (câu)	Khối lượng (KB)
1	Từ điển Lạc Việt	Tiếng Việt	5789	347
2	Báo VOV Online	Tiếng Việt	2778	1363
3	EuroParl Corpus [5]	Tiếng Pháp	35728	2145
4	BTEC Corpus [5]	Tiếng Anh	39821	2387
5	DATIC Corpus	Tiếng Pháp	11829	798
6	Các Corpus khác	Tiếng Anh	4892	293
7	Các nguồn khác	Tiếng Việt	5597	356
Tổng cộng			106434	7689

Giải pháp [4] đã thực hiện việc đa ngữ hoá khoa dữ liệu bằng cách sử dụng các công cụ dịch tự động (chủ yếu là Google Translator) để dịch trực tiếp từ tiếng Việt sang tiếng Anh, tiếng Pháp; từ tiếng Anh sang tiếng Việt, tiếng Pháp; từ tiếp Pháp sang tiếng Anh, tiếng Việt.

Dựa trên nguồn dữ liệu này, chúng tôi tiến hành việc thử nghiệm giải pháp đa ngữ hoá do chúng tôi để xuất để có thể khai thác ở nhiều ngôn ngữ khác nhau.

III. ĐỀ XUẤT

Hầu hết các kho dữ liệu đa ngữ hiện nay được xây dựng bằng cách sưu tập dữ liệu đơn ngữ cho từng ngôn ngữ hoặc sưu tập từ một ngôn ngữ và sau đó dịch sang ngôn ngữ khác để có được dữ liệu ở các ngôn ngữ mong muốn. Điều này chẳng những làm tăng chi phí xây dựng cho các cặp dữ liệu mà còn sẽ gặp nhiều khó khăn đối với những cặp ngôn ngữ thiếu tài nguyên hoặc không tương đồng cấu trúc ngữ pháp. Trong bài báo này chúng tôi đề xuất giải pháp sử dụng một ngôn ngữ làm ngôn ngữ trục để xây dựng, mô hình tổng quát như sau:



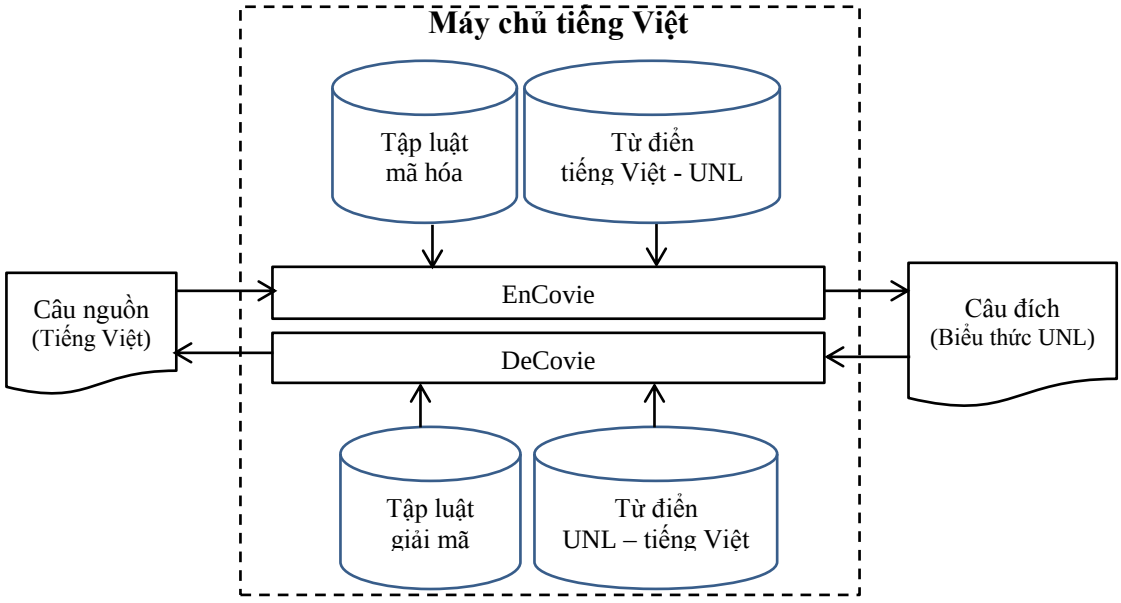
Hình 3. Mô hình tổng quát xây dựng kho dữ liệu đa ngữ

Trong mô hình đề xuất này, khi xây dựng các kho dữ liệu đa ngữ, chúng ta chỉ cần sưu tập dữ liệu ở một ngôn ngữ nào đó (đã được hỗ trợ bởi hệ thống UNL), sau đó sử dụng phần mềm EnConverter để chuyển dữ liệu này sang ngôn ngữ UNL và cuối cùng dịch dữ liệu từ UNL sang các ngôn ngữ khác.

Mấu chốt của mô hình này là mỗi ngôn ngữ cần phải nghiên cứu xây dựng phần mềm EnConverter và DeConverter để thực hiện quá trình dịch tự động.

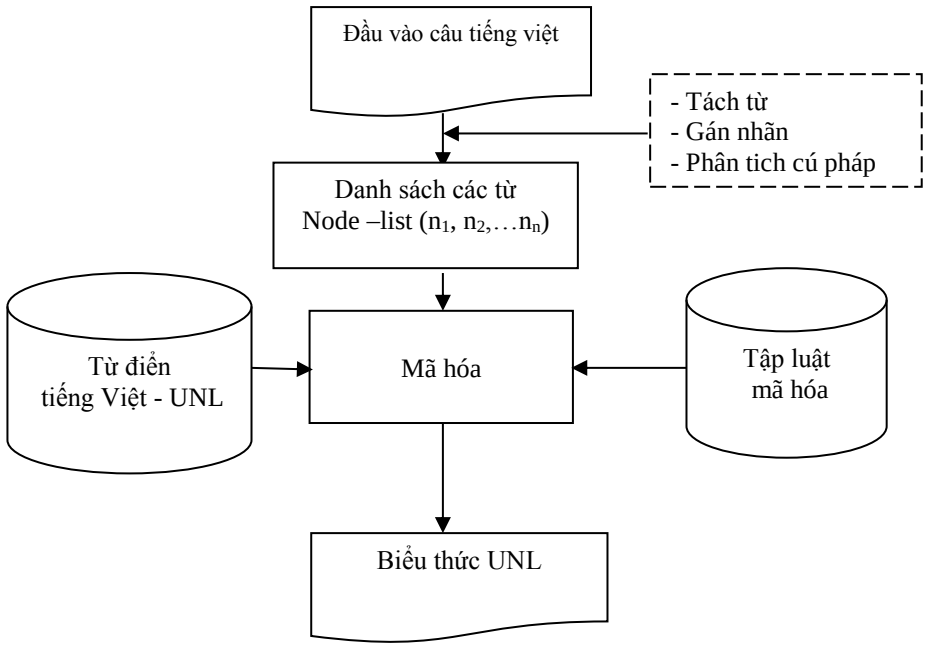
IV. KẾT QUẢ VÀ BÌNH LUẬN

Chúng tôi đã xây dựng máy chủ tiếng Việt gồm hai chức năng là chuyển đổi văn bản được viết trong ngôn ngữ Việt sang văn bản được viết trong ngôn ngữ UNL được thực hiện bởi công cụ gọi là EnCovie và ngược lại chuyển đổi văn bản được viết trong ngôn ngữ UNL sang văn bản được viết trong tiếng Việt được thực hiện bởi công cụ DeCovie.



Hình 4. Mô hình máy chủ tiếng Việt

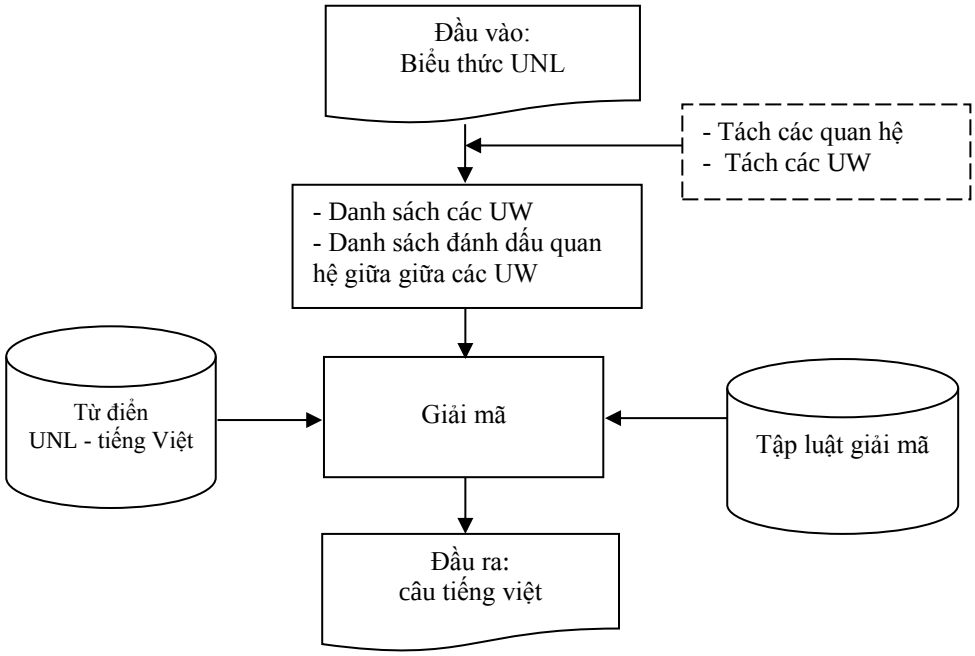
EnCovie là công cụ dùng để chuyển đổi một câu tiếng Việt sang một biểu thức UNL. Công cụ này có khả năng phân tích độc lập về hình thái, cú pháp và ngữ nghĩa.



Hình 5. Sơ đồ chuyển đổi của công cụ EnCovie

Với câu đầu vào tiếng Việt, chúng tôi sẽ được thực hiện tách từ, gán nhãn từ loại và phân tích cú pháp câu tiếng Việt bởi một môđun tích hợp. Kết quả của giai đoạn tiền xử lý này là các từ, các từ sau khi tách ra được lưu trữ trên các nút (n_1,n_2,...,n_n) của danh sách gọi là Node-list. Trong Node-list, nút đầu tiên của danh sách gọi là nút "head" và nút cuối cùng của danh sách gọi là nút "last". Tiếp theo tìm và tạo liên kết giữa các nút trong Node-list với các mục từ trong từ điển tiếng Việt - UNL. Chỉ cần tìm headword và thuộc tính của nút so sánh với mục từ trong từ điển, nếu được tìm thấy thì hệ thống sẽ lưu thông tin và tạo liên kết nhưng nếu ngược lại sẽ sử dụng chính từ gốc làm UW và gắn thuộc tính bằng "null" và đồng thời chèn từ mới này vào từ điển tiếng Việt - UNL. Công cụ Encovie sẽ quét từ trái sang phải các nút trên Node-list thông qua hai cửa sổ phân tích trái (LW) và cửa sổ phân tích phải (RW). EnCovie sử dụng hai cửa sổ LW và RW kiểm tra điều kiện thõa của hai nút liền kề để thực hiện luật. Quá trình mã hóa kết thúc khi sửa sổ LW đi đến hết Node-list và tạo ra một biểu thức UNL biểu diễn tương đương với câu tiếng Việt đầu vào.

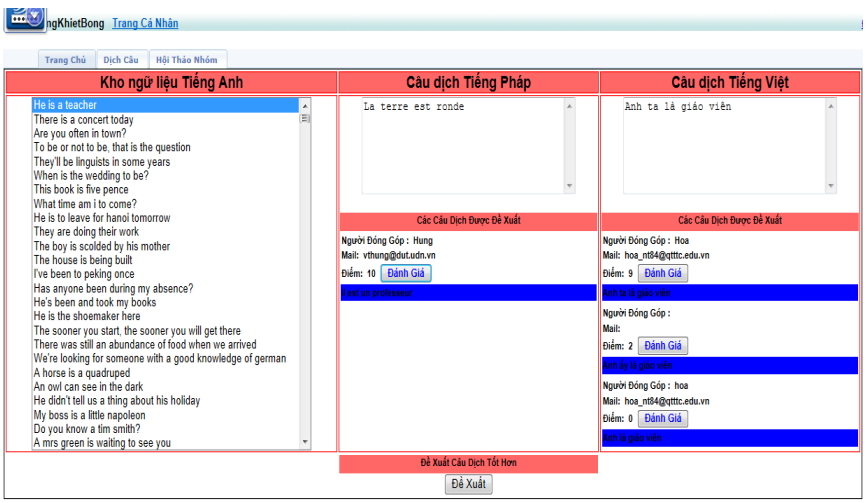
Để chuyển đổi biểu thứ UNL sang câu tiếng Việt tương ứng, chúng tôi xây dựng công cụ DeCovie được mô tả như Hình 6.



Hình 6. Sơ đồ chuyển đổi của công cụ DeConverter

Với biểu thức UNL đầu vào là tập hợp các mối quan hệ nhị phân, một môđun sẽ phá vỡ các mối quan hệ nhị phân đó bằng cách thực hiện việc tách các quan hệ và UWs. Mỗi UW sẽ được lưu trữ thành danh sách các nút Node (n_1,n_2,...n_n), mỗi quan hệ được lưu trên danh sách gọi là Rel (r_1,r_2,...r_n) và có một liên kết giữa các UW và quan hệ. Trên cơ sở phân tích ngữ nghĩa và từ vựng, DeCovie sẽ chuyển biểu thức UNL sang một câu tiếng Việt. Thông tin chi tiết về việc phát triển và thử nghiệm máy chủ tiếng Việt để dịch tự động giữa tiếng Việt - UNL được trình bày kỹ trong luận án tiến sĩ của Phan Thị Lệ Thuyền [6].

Bên cạnh đó, chúng tôi đã xây dựng môi trường để hỗ trợ người dùng chỉnh sửa dữ liệu đa ngữ:



Hình 7. Hiệu chỉnh dữ liệu đa ngữ

Với công cụ hỗ trợ này, người dùng có thể chọn ngôn ngữ, chọn câu trong kho dữ liệu cần chỉnh sửa và ghi vào câu dịch đúng (theo ý của người sửa) và hệ thống sẽ ghi lại kết quả (có thể có nhiều phiên bản cho một câu, trong một ngôn ngữ nào đó).

Với cách làm như vậy, chúng tôi đã thử nghiệm sử dụng UNL để tạo kho dữ liệu được mô tả ở Bảng 1 sang 03 ngôn ngữ Anh, Pháp và Việt.

Để so sánh kết quả đạt được từ hai phương pháp khác nhau (dịch trực tiếp bằng Google Translator và dịch qua UNL), chúng tôi sử dụng các phương pháp NIST (National Institute of Standards and Technology) và BLEU (BiLingual Evaluation Understudy) để đánh giá [7]. Kết quả như sau:

Bảng 2. Bảng so sánh kết quả NIST và BLEU

STT	Nguồn gốc	Google Translator		UNL	
		NIST	BLEU	NIST	BLEU
1	Từ điển Lạc Việt	7.235	0,114	7.835	0,080
2	Báo VOV Online	6.512	0,321	7.124	0,118
3	EuroParl Corpus	7.437	0,102	7.916	0,076
4	BTEC Corpus	7.321	0,101	7.623	0,084
5	DATIC Corpus	6.815	0,352	7.451	0,102
6	Các Corpus khác	6.762	0,334	7.323	0,121
7	Các nguồn khác	7.016	0,112	7.563	0,090

Qua kết quả ở Bảng 2, chúng tôi nhận thấy với các dữ liệu song ngữ và khá chuẩn như các câu song ngữ trích từ Từ điển Lạc Việt, từ các kho dữ liệu song ngữ EuroParl Corpus, BTEC Corpus thì kết quả dịch qua các phần mềm dịch Google Translator và UNL tốt hơn so với các dữ liệu trích xuất từ báo VOV hoặc các dữ liệu khác.

Một vấn đề khó khăn nhất trong giải pháp này là việc kiểm tra và hiệu chỉnh dữ liệu UNL là rất khó và đòi hỏi phải có chuyên gia nghiên cứu về UNL mới chỉnh sửa được (giống như chúng ta phải biết tiếng Việt, tiếng Anh để chỉnh sửa các câu trong ngôn ngữ tương ứng).

V. KẾT LUẬN

Chúng tôi đã xây dựng kho dữ liệu đa ngữ nhờ vào hệ thống dịch tự động UNL và môi trường hợp tác. Kết quả dịch thông qua UNL được so sánh với kết quả trước đó được dịch bằng Google Translator và chất lượng đạt được tương đương nhau. Trên cơ sở kết quả đạt được, chúng tôi sẽ tiếp tục thực hiện việc thu thập dữ liệu để làm phong phú thêm kho dữ liệu và tiến hành dịch ra nhiều ngôn ngữ khác như Nga, Trung, Nhật, Hàn, Đức,... đặc biệt là tiếng dân tộc của Việt Nam như tiếng Chăm, tiếng Khmer,…

Hiện nay nhiều hệ thống dịch đa ngữ được xây dựng với sự hỗ trợ của Deep Learning đã nâng độ chính xác của hệ thống dịch, tuy nhiên để có kết quả như vậy thì chúng ta cũng cần có nguồn dữ liệu khá lớn. Nghiên cứu của chúng tôi sẽ là giải pháp lựa chọn đối với những ngôn ngữ nghèo nàn tài nguyên và giảm chi phí xây dựng cặp ngôn ngữ. Chúng tôi sẽ ứng dụng kết quả này vào một số mục đích như: dạy và học ngoại ngữ, tiếp tục cập nhật và bổ sung một cách tự động để mở rộng kho dữ liệu nhằm phục vụ cho nhu cầu dịch tự động cũng như phát triển hoàn thiện hệ thống dịch tự động theo mô hình dịch thống kê.

TÀI LIỆU THAM KHẢO

[1] Võ Trung Hùng, *Một số phương pháp và mô hình áp dụng trong xử lý ngôn ngữ tự nhiên*, Nhà xuất bản Thông tin và Truyền thông, ISBN: 987-604-80-2414-7, 2017.

[2] UNL Center, The UNL Specifications, available at http://www.undl.org (accessed 13 July 2023), 2004.

[3] Hiroshi Uchida, Meiying Zhu, *Tarcisio Della Senta, Universal Networking Language*, Published by UNDL Foundation, ISBN 2-8399-0128-5/978-2-8399-0128-4, 2005.

[4] Nguyễn Thị Hoa, Võ Trung Hùng, "Xây dựng kho dữ liệu đa ngữ Anh - Pháp - Việt," *Tạp chí Khoa học Công nghệ Đại học Đà Nẵng*, số 10 (59), pp. 50-57, 2012.

[5] Philipp Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," *Proceedings of Machine Translation Summit X: Papers*, 2005.

[6] Phan Thị Lệ Thuyền, "Sử dụng ngôn ngữ trục trong dịch đa ngữ," Luận án tiến sĩ, Đại học Đà Nẵng, 2018.

[7] Võ Trung Hùng, "Phương pháp và công cụ đánh giá tự động các hệ thống dịch tự động trên mạng," *Tạp chí Khoa học và Công nghệ, Đại học Đà Nẵng*, số: 1 (18), tr. 37-42, 2007.

BUILDING A MULTILANGUAL CORPUS WITH UNL

Vo Trung Hung, Phan Thi Le Thuyen, Ninh Khanh Chi

ABSTRACT: *This article presents the construction of a multilingual data warehouse using a language called UNL (Universal Networking Language). UNL is defined as consisting of components corresponding to a natural language and it can represent all the information of the language in a structured form without causing semantic ambiguity. With this feature, the UNL language can be used as an intermediate language (pivot language) to convert between natural languages instead of directly translating between two languages that we often use. Using the UNL language in a multilingual translation system has two major advantages: UNL has no semantic ambiguity and reduces the number of translation pairs from n*(n-1)/2 to 2*n. We tested with 106,434 sentences, the output results showed that translation via UNL is better than direct translation using Google Translator (the evaluation is based on two methods NIST and BLEU).*