Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00011

CHUYỂN ĐỔI CÂU ĐƠN TIẾNG VIỆT SANG BIỂU THỨC UNL

Phan Thị Lệ Thuyền [1], Võ Trung Hùng[2]
[1] Trường Đại học Quang Trung
[2] Đại học Đà Nẵng
thuyenptl@gmail.com, vthung@dut.udn.vn

TÓM TẮT— *UNL (Universal Networking Language) là một ngôn ngữ nhân tạo và có thể diễn đạt các nội dung trong ngôn ngữ tự nhiên theo cách thức của nó. UNL là cơ sở để phát triển các phần mềm dịch tự động đa ngữ thông qua ngôn ngữ trục (trường hợp này là ngôn ngữ UNL). UNL mở ra khả người sử dụng có thể truy cập thông tin trên mạng Internet mà không bị rào cản về ngôn ngữ. UNL đã được nghiên cứu và ứng dụng cho 48 ngôn ngữ khác nhau. Hệ thống UNL bao gồm hai thành phần chính là mã hóa (EnCoverter) và giải mã (DeConverter). Mã hóa là quá trình chuyển đổi một văn bản từ ngôn ngữ nguồn (ví dụ tiếng Anh, tiếng Việt,...) sang một văn bản viết bằng ngôn ngữ UNL và giải mã là quá trình chuyển đổi ngược lại (từ văn bản viết trong ngôn ngữ UNL sang ngôn ngữ đích). Hiện nay, hệ thống UNL cho tiếng Việt chưa được phát triển. Trong bài báo này, chúng tôi trình bày kết quả nghiên cứu về phương pháp mã hóa một câu đơn tiếng Việt sang biểu thức UNL tương ứng. Để thực hiện việc chuyển đổi, chúng ta phải xây dựng từ điển Việt – UNL, các luật (quy tắc ngữ pháp) và phần mềm để chuyển đổi. Đặc biệt, chúng tôi đã đề xuất một thuật toán sử dụng các luật trong tập luật mã hóa để tạo ra các thuộc tính tương ứng của UNL và giải quyết các mối quan hệ khác khi chuyển đổi. Kết quả đạt được là chúng tôi đã xây dựng công cụ EnCoVie thực hiện chức năng mã hóa cho một số trường hợp của câu đơn tiếng Việt.*

Từ khóa— *Dịch máy, ngôn ngữ mạng dùng chung, dịch liên ngôn ngữ, xử lý ngôn ngữ tự nhiên, mã hóa.*

I. GIỚI THIỆU

Hiện nay, có nhiều hệ thống dịch tự động đa ngữ trực tuyến miễn phí như Google Translate, Systran Machine Translation, Reverso Translator,... nhưng chất lượng bản dịch vẫn còn là một vấn đề lớn [1][2]. Các hệ thống dịch đa ngữ hiện nay đang xây dựng theo hai hướng là dịch trực tiếp giữa các ngôn ngữ với nhau hoặc dịch thông qua một ngôn ngữ trung gian (lấy một ngôn ngữ làm trung gian, ví dụ như tiếng Anh, để dịch chuyển tiếp sang ngôn ngữ khác). Tuy nhiên, dịch qua ngôn ngữ trung gian kết quả không thể tốt bằng dịch trực tiếp [3]. Tuy nhiên, để dịch cho từng cặp ngôn ngữ trực tiếp thì số lượng các phần mềm dịch là rất lớn (nếu có n ngôn ngữ chúng ta cần đến n*(n-1)/2 cặp dịch). Mặt khác, để dịch trực tiếp cho mỗi cặp ngôn ngữ, chúng ta phải tiến hành nghiên cứu về từ vựng, cú pháp, ngữ nghĩa và gặp nhiều khó khăn như sự khác biệt quá xa giữa các ngôn ngữ hoặc thiếu các nguồn tài nguyên phục vụ việc dịch (từ điển, quy tắc ngữ pháp,...) [4]. Trong khi đó dịch thông qua một ngôn ngữ trung gian chỉ cần 2*n cặp dịch và hạn chế những khó khăn cho những cặp ngôn ngữ hạn chế về thiếu nguồn tài nguyên. Dịch thông qua một ngôn ngữ trung gian không phải là ngôn ngữ tự nhiên là một hướng nghiên cứu đang được quan tâm khi phát triển các hệ thống dịch đa ngữ. Nó tạo ra khả năng tích hợp các hệ thống dịch riêng lẻ lại với nhau và giảm chi phí xây dựng các cặp dịch trực tiếp [3].

Một trong những hệ thống hỗ trợ đa ngữ hóa và dịch tự động được nghiên cứu hiện nay là UNL. Mục đích chính của UNL là cung cấp cho người sử dụng Internet khả năng truy cập vào các trang web bằng ngôn ngữ của họ [3]. Cộng đồng các nhà nghiên cứu về dịch tự động Universal Networking Digital Language (UNDL) đã cung cấp hai công cụ EnCo và DeCo để thực hiện chức năng mã hóa từ một ngôn ngữ tự nhiên sang biểu thức UNL và giải mã từ một biểu thức UNL sang ngôn ngữ tự nhiên. Các hệ thống ứng dụng UNL thực hiện chức năng chuyển đổi ngôn ngữ tự nhiên sang biểu thức UNL đã được xây dựng như: hệ thống IAN (http://www.unlweb.net) được phát triển trên nền tảng web để chuyển đổi một ngôn ngữ tự nhiên sang UNL, hệ thống mã hóa tiếng Punjabi của Parteek Kumar [6], mã hóa tiếng Anh của Manoj Jain and Om P. Damani [7], mã hóa tiếng Ta-min của J Balaji [8], mã hóa tiếng Bangla của Md. Nawab Yousuf Ali [9].

Vấn đề đặt ra là làm thế nào để tích hợp tiếng Việt vào nền tảng UNL để tạo thành một hệ thống đa ngữ trong khi chưa có một nghiên cứu chính thức nào cho tiếng Việt? Hệ thống này bao gồm hai chức năng chính là thực hiện mã hóa từ tiếng Việt sang UNL và giải mã từ UNL sang tiếng Việt. Nếu làm được việc đó, chúng ta có thể dịch từ tiếng Việt sang bất cứ ngôn ngữ nào đã tích hợp vào UNL và ngược lại.

Trong bài báo này, chúng tôi đề xuất một hướng tiếp cận mới về vấn đề dịch tự động cho tiếng Việt dựa vào UNL. Chúng tôi thử nghiệm trước hết cho các câu đơn tiếng Việt. Để làm việc này, trước hết chúng tôi xác định sự tương đương giữa các từ loại, thuộc tính trong câu tiếng Việt với các thuộc tính UW (Universal Word) của UNL; tiếp đến là xử lý các quan hệ giữa các UWs trong biểu thức UNL. Trên cơ sở đó, chúng tôi đề xuất thuật toán tạo ra một biểu thức UNL từ một câu đơn tiếng Việt tương ứng mà trọng tâm là giải quyết các vấn đề liên quan đến thuộc tính và quan hệ của các UWs.

Bài báo được tổ chức thành các phần chính như sau: sau phần giới thiệu là phần trình bày những kết quả nghiên cứu liên quan; phần thứ ba giới thiệu đề xuất của chúng tôi để áp dụng UNL cho tiếng Việt và quá trình mã hóa một câu tiếng Việt trong hệ thống; phần thứ tư trình bày kết quả thử nghiệm và đánh giá; cuối cùng là phần kết luận nhằm trình bày kết quả đạt được và hướng phát triển.

II. CÁC NGHIÊN CỨU LIÊN QUAN

A. Cấu trúc biểu thức UNL

UNL là một ngôn ngữ giả có khả năng mô phỏng thế giới ngôn ngữ tự nhiên. Kết quả là nó cho phép người sử dụng có thể biểu diễn tất cả các trí thức từ ngôn ngữ 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, nút biểu diễn khái niệm và cạnh biểu diễn mối quan hệ giữa các khái niệm [10].

Một biểu thức UNL được xem như một câu trong ngôn ngữ tự nhiên. Nó được tạo nên từ ba yếu tố chính: các từ vựng, các thuộc tính và các quan hệ. Các từ vựng được liên kết với nhau nhờ các quan hệ để tạo thành một biểu thức UNL tương ứng với một câu của ngôn ngữ tự nhiên. Các thuộc tính mô tả các thông tin chủ quan, thể hiện quan điểm của người nói được diễn đạt [3].

Ví dụ một câu đầu vào tiếng Anh "john will have finished his project" được chuyển sang UNL tương đương như sau:

```
{unl}
agt(finish(icl>act>do,equ>land_up,agt>person,gol>thing).@entry.@future.
    @complete,john(icl>name>abstract_thing,com>male,nam<person))
pos(project(icl>labour>abstract_thing,pos>thing,pur>uw),he(icl>person))
gol(finish(icl>act>do,equ>land_up,agt>person,gol>thing).@entry.@future.
    @complete,project(icl>labour>abstract_thing,pos>thing,pur>uw))
{/unl}
```

Trong đó:

- "agt", "pos" và "gol" là các quan hệ.
- "finish(icl>act>do,equ>land_up,agt>person,gol>thing)", "john(icl>name>abstract_thing,com>male,nam<person)", "project(icl>labour>abstract_thing,pos>thing,pur>uw)", "he(icl>person)" và "finish(icl>act>do,equ>land_up,agt>person,gol>thing)" là các từ vựng.
- "@entry","@future","@complete" là các thuộc tính.

B. Định dạng luật mã hóa

Chúng tôi xây dựng công cụ EnCoVie với định dạng luật mã hóa được thiết kế dựa trên nguyên tắc của UNL EnConverter Specifications [3] như sau:

```
    ký hiệu luật {COND1:ACTION1:REL1}{COND2:ACTION2:REL2};
```

Trong đó,

- <COND1> và <COND2> chỉ điều kiện 1 và 2, chứa các thuộc tính từ vựng và ngữ nghĩa của cửa sổ phân tích trái và phải.
- <ACTION1> và <ACTION2> chỉ hành động được thực hiện nếu điều kiện tương ứng đúng.
- <REL1> và <REL2> chỉ ra mối quan hệ có thể có giữa hai cửa sổ phân tích.

Ví dụ: ta có luật >{N: null: aoj}{ADJ:+R:null};

Đây là luật sửa đổi phải (>), kết quả là xoá nút trái từ danh sách các nút. COND1 là một danh từ, COND2 là tính từ. ACTION1 chứa "null" nên không cần phải làm gì trên các cửa sổ phân tích trái, ACTION2 thêm thuộc tính "R" vào cửa sổ phải. REL1 chứa "aoj" là tạo mối quan hệ AOJ giữa hai cửa sổ, REL2 chứa "null" là không có quan hệ.

C. Cấu trúc từ điển tiếng Việt - UNL

Một mục từ của từ điển bao gồm ba phần cơ bản dựa trên EnConverter Specifications [3]: một headword (từ đầu mục từ), một từ vựng và một tập các thuộc tính ngữ pháp. Định dạng dữ liệu cho các mục từ trong từ điển tiếng Việt – UNL như sau:

```
    [HW]"UW"(ATTR,ATTR,…)<FLG,FRE, PRI>;
```

Trong đó,

HW: từ đầu mục từ của ngôn ngữ;

UW: từ vựng;

ATTR: thuộc tính ngữ pháp;

FLG: cờ ngôn ngữ;

FRE: tần số xuất hiện;

PRI: mức ưu tiên.

D. Các mô hình câu đơn trong tiếng Việt

Câu đơn là câu có một kết cấu chủ - vị, nghĩa là kết cấu có hai vế được đặt theo quan hệ cú pháp cơ bản là quan hệ chủ ngữ và vị ngữ. Ví dụ câu "cô ấy thức suốt đêm", "cô ấy" đóng vai trò là chủ ngữ của câu và "thức suốt đêm" là vị ngữ trong câu với "suốt đêm" là bổ ngữ cho động từ "thức". Theo [10], trong tiếng Việt các câu đơn có thể quy thành 12 mô hình tiêu biểu để biểu hiện các phạm trù ý nghĩa khác nhau: xác định, liên hệ, quá trình, hành động, đặc trưng, tồn tại,... Khi nghiên cứu về câu đơn, chúng tôi thấy rằng có sự biến thể cấu trúc câu và sự phức tạp của câu không phải do ý nghĩa từ vựng riêng lẻ của các yếu tố trong câu mà do cấu trúc ngữ nghĩa của cả câu quy định. Quá trình phức tạp hóa câu đơn thường thường bao gồm nhiều tầng, nhiều lớp, nghĩa là trong mỗi kết cấu chủ - vị lại mang thêm kết cấu chủ - vị khác, hoặc thêm thành phần chủ ngữ, vị ngữ, bổ ngữ, định tố khác và cứ thế mở rộng theo mức độ lỏng – chặt khác nhau.

Các mô hình tiêu biểu câu đơn

STT	Kiểu mô hình	Kiểu vị ngữ	Ví dụ minh họa
1	1 từ	Không có vị ngữ, không có chủ ngữ	Chào!
2	0 – V	Vắng chủ ngữ, vị ngữ động từ	Sắp sang xuân!
3	Ø – V	Zêro chủ ngữ, vị ngữ động từ	Cháy nhà!
4	C – V	Vị ngữ là: "là"+(danh từ, tính từ, động từ) và có biến thể không có hệ từ "là"	Cô ấy là sinh viên
5	C – V	Vị ngữ là động từ nội động	Tôi làm việc
6	C – V – B	Vị ngữ là động từ ngoại động	Tôi đến để gặp anh ta
7	C1 – V1 – C2 - V2	Vị ngữ là động từ sai khiến	Hoa bắt các em ăn
8	C – V1 – V2 – B	Vị ngữ là động từ + động từ, với B là bổ ngữ	Học sinh yêu cầu giải đáp thắc mắc
9	C.Vp.V.B	Câu bị động, với Vp là động từ làm vị ngữ	Tôi được khen
10	C – V	Vị ngữ là danh từ + tính từ	Mẹ tôi tính tình hiền lành
11	C – V	Vị ngữ là thành ngữ, quán ngữ	Thằng này mặt người dạ thú
12	Cx – Vx – Bx	Câu đơn khai triển	Tôi gọi nó đọc bài

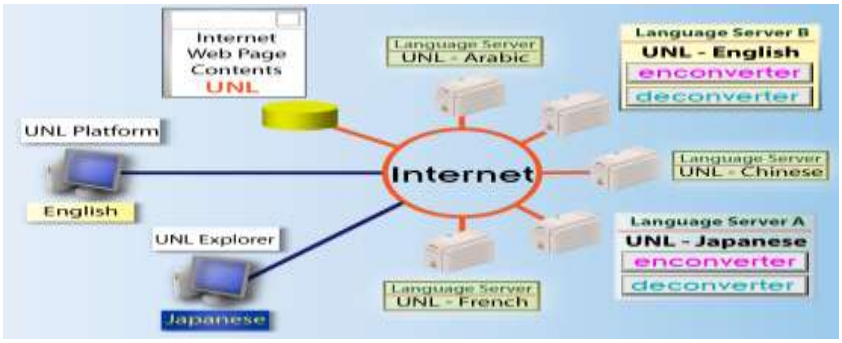
III. GIẢI PHÁP ĐỀ XUẤT

A. Hệ thống UNL cho tiếng Việt

Hệ thống dịch tự động đa ngữ UNL có thể bao gồm nhiều máy chủ ngôn ngữ khác nhau cho tiếng Anh, tiếng Việt, tiếng Pháp,... Mỗi máy chủ ngôn ngữ sẽ đảm nhận 2 chức năng đó là dịch một văn bản từ ngôn ngữ này sang ngôn ngữ UNL (mã hóa) và dịch ngược lại (giải mã). Ví dụ, người sử dụng muốn dịch một văn bản từ tiếng Việt sang tiếng Anh thì văn bản tiếng Việt sẽ được gửi đến máy chủ tiếng Việt để dịch từ tiếng Việt sang UNL, sau đó văn bản UNL sẽ gửi sang máy chủ tiếng Anh để dịch từ UNL sang tiếng Anh và kết quả được trả về cho người sử dụng.

Các máy chủ ngôn ngữ có thể 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. Chúng ta cũng có thể đăng ký với tổ chức Universal Networking Language Foundation (http://www.undlfoundation.org/undlfoundation/) để tích hợp lên máy chủ chung của UNL. Hiện tại, chúng tôi đang ở bước nghiên cứu và thử nghiệm cho tiếng Việt nên đang cài đặt máy chủ tiếng Việt riêng mà chưa tích hợp lên máy chủ UNL.

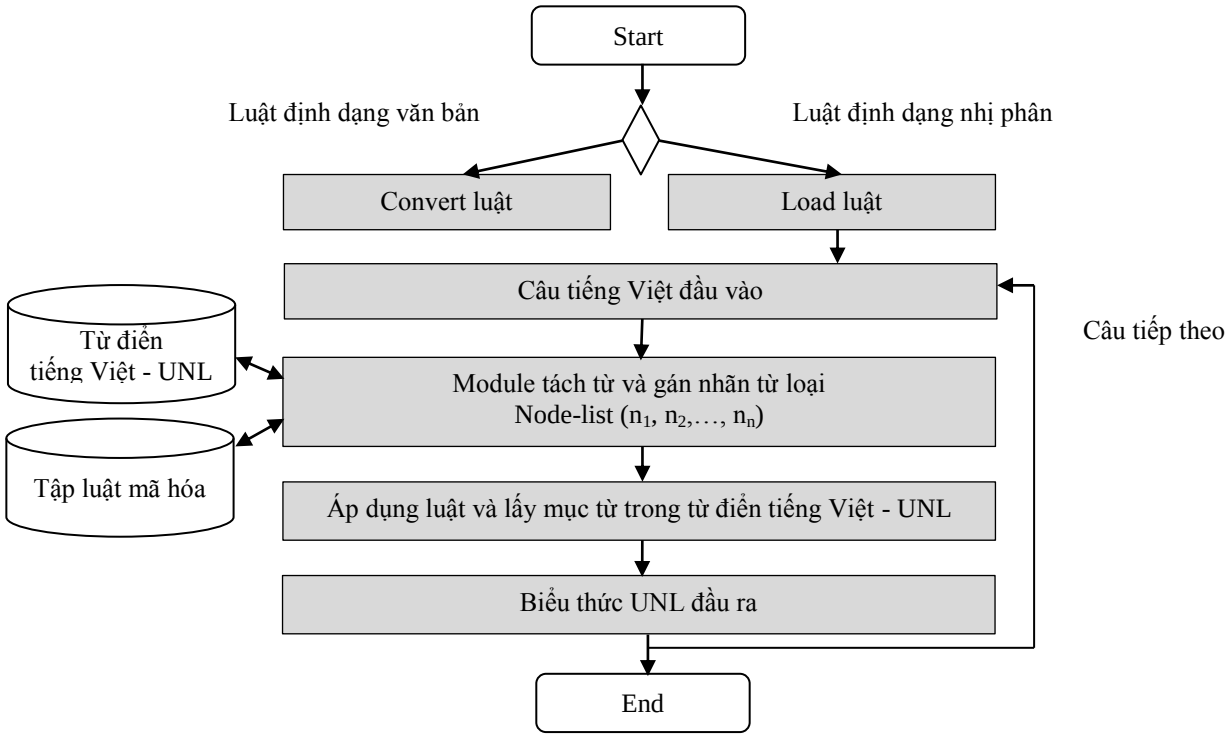
Mô hình của hệ thống như sau:



Hình 1. Hệ thống UNL cho các ngôn ngữ

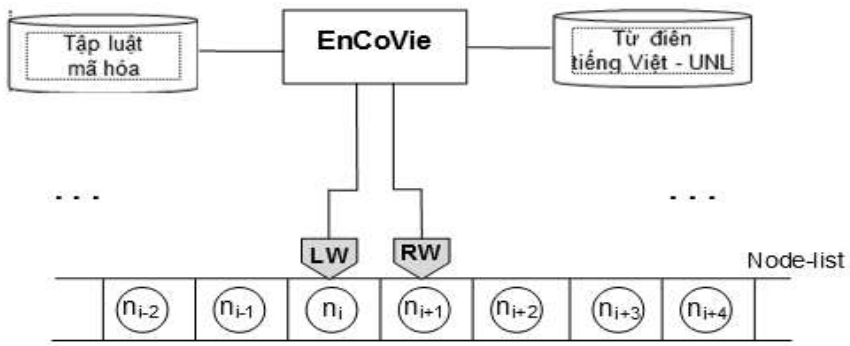
B. Mã hóa câu tiếng Việt

Sơ đồ chuyển đổi câu tiếng Việt sang biểu thức UNL như sau:



Hình 2. Sơ đồ chuyển đổi câu tiếng Việt sang biểu thức UNL

Quá trình chuyển đổi một câu tiếng Việt sang biểu thức UNL như sau. Trước hết, thực hiện việc tách từ trong câu đầu vào và gán nhãn cho mỗi từ. Mỗi từ của câu đầu vào sẽ được lưu trữ trên các nút (n1,n2,...nn) 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 công cụ EnCoVie sẽ tìm và thực hiện mối liên kết giữa các nút trong Node-list với các mục từ trong từ điển tiếng Việt – UNL, nếu không có mục từ tương đồng thì công cụ sẽ ưu tiên mục từ gần nghĩa từ loại nhất (ví dụ hai định nghĩa khái niệm: [đường] “sugar(icl>sweetening>thing)”(N), [đường]”street(icl>thoroughfare> thing)”(N) thì hệ thống sẽ xem xét các từ lân cận ở phía trước và phía sau để đưa ra lựa chọn). Công cụ EnCoVie sẽ duyệt qua các nút của Node-list, câu đầu vào sẽ được quét từ trái sang phải thông qua hai cửa sổ phân tích trái (LW) và cửa sổ phân tích phải (RW). LW và RW kiểm tra hai nút liền nhau có thỏa mãn các điều kiện của một trong các luật trong tập luật để thực hiện mã hóa. Quá trình mã hóa bắt đầu LW nằm ở nút head và kết thúc của trình khi RW nằm nút last.



Hình 3. Mô hình việc mã hóa cho tiếng Việt

Để tạo biểu thức UNL, vấn đề chính là cần phải xác định mối quan hệ giữa các UWs và bổ sung các thuộc tính cho UW, chúng tôi đề xuất các bước để giải quyết hai vấn đề trên như sau:

Bước 1. Trong danh sách các nút trên Node-list. LW nằm ở nút đầu (head) của Node-list và nút kế tiếp sẽ là RW.

Bước 2. Công cụ EnCoVie sẽ bắt đầu tìm luật mã hóa từ tập luật.

Bước 3. Xét điều kiện luật:

- Nếu thỏa mãn các điều kiện của LW và RW thì sẽ thực hiện luật (sửa đổi thuộc tính cho UW hoặc tạo mối quan hệ giữa các UWs). Sau khi thực hiện xong, LW và RW dịch chuyển sang trái một nút và chuyển sang bước 2.

- Nếu không tìm thấy luật phù hợp nào trong tập luật thì thực hiện di chuyển LW và RW chuyển sang phải một nút và chuyển sang bước 4.

Bước 4. Kiểm tra trong danh sách Node-list:

- Nếu chỉ có một nút duy nhất (trừ nút head và nút last) thì nút này sẽ là nút vào (entry) và quá trình mã hóa đã kết thúc.
- Nếu không phải là nút duy nhất thì chuyển sang bước 2.

Giải thuật cho thuật toán xác định mối quan hệ và các thuộc tính như sau:

```
Dữ liệu vào: Node-list{n1,n2,…nn}
Begin
LW ={n0};
RW={n1};
While RW={Nn+1} do
If ({một luật được tìm thấy})
  {
   If (luật (:))
    sửa đổi thuộc tính của nút;
   Else
      If (luật (+)hoặc (-))
        {
         Tạo một nút kết hợp;
         Xóa nút;
         LW → trái;
         RW→ trái;
         }
       Else
         {
          Tạo mối quan hệ giữa hai UW;
          Xóa nút;
          LW → trái;
          RW→ trái;
         }
      Endif
  Endif
  }
Else
 {
 LW → phải;
 RW→ phải;
}
Endif
Endwhile
If i=1 then
   Nút ni= "+.@entry"
Endif
End.
```

IV. THỬ NGHIỆM VÀ ĐÁNH GIÁ

A. Thử nghiệm

1. Dữ liệu thử nghiệm

Trong bài báo này, chúng tôi chọn phân tích trên mô hình thứ 4 của 12 mô hình để thử nghiệm vì đây là mô hình có kết cấu Chủ - Vị như là cấu trúc câu cơ bản của dạng thức câu đơn. Mô hình thứ 4, chúng tôi không giải quyết trường hợp C- V (với V là động từ) vì trường hợp V là động từ thì khi đó biến thể của nó là các mô hình 5, 6, 7, 8, 9:

a. Câu hai thành phần có vị ngữ là danh từ

$$\mathrm{C} - \frac{V}{(\text{là} + danh\ \text{từ})}$$

Ví dụ: Anh ấy là chiến sĩ thi đua.

Mô hình này có những biến thể khác nhau tùy thuộc vào vị ngữ:

- Đại từ + hệ từ "là" + tổ hợp danh từ có giới từ
  Ví dụ: Đó là một phát hiện quan trọng đối với hòa bình.
- Danh từ + động từ (hệ từ) + danh từ;
  Ví dụ: Tôi sắp trở thành sinh viên.

b. Câu hai thành phần có vị ngữ là tính từ

$$\mathrm{C} - \frac{V}{(\text{là} + tính\ \text{từ})}$$

Ví dụ: Hồng là rất đẹp.

Mô hình này vị ngữ cũng có thể kết hợp với động từ (hóa ra, trở nên,...). Ví dụ: Ba ta hóa ra khó chịu.

c. Câu hai thành phần có vị ngữ là danh từ hoặc tổ hợp danh từ không có hệ từ "là"

$$\mathrm{C} - \frac{V}{(danh\ \text{từ})}$$

Ví dụ: Cả nước một lòng.

Mô hình này có khả năng chấp nhận những biến thể sau đây:

- Chủ ngữ + số từ + danh từ
  Ví dụ: Điện cao thế ba pha.
- Chủ ngữ + từ so sánh + danh từ
  Ví dụ: Thân em như tấm lụa đào.
- Chủ ngữ + đại từ
  Ví dụ: ông ấy.
- Chủ ngữ + loại từ + danh từ
  Ví dụ: mỗi người một phòng.

Chúng tôi đã xây dựng tập câu tiếng Việt tuân thủ theo 3 cấu trúc và 4 biến thể trên với nhiều loại từ làm chủ ngữ để làm dữ liệu thử nghiệm. Tuy chúng tôi xây dựng trên 7 mẫu câu đơn nhưng lại giải quyết thêm các trường hợp với chủ ngữ có từ loại khác nhau. Ví dụ câu a. Ta có: " C + "là"+danh từ", chủ ngữ có thể là danh từ, đại từ thì ta có thể xây dựng tập luật mã hóa cho cấu trúc "Danh từ + "là"+ danh từ" và "đại từ + "là"+ danh từ".

2. Công cụ mã hóa

Để chuyển đổi một câu tiếng Việt sang biểu thức UNL, chúng tôi xây dựng công cụ với chức năng mã hóa gọi là EnCoVie. Để thực hiện chức năng tách từ và gán nhãn từ loại, chúng tôi xây dựng một module để xử lý và module này chúng tôi kế thừa các tài nguyên và công cụ VLSP (http://vlsp.vietlp.org:8080/demo/?page=resources). Chúng tôi sử dụng bộ từ điển tiếng Việt – UNL của nhóm tác giả [9][10] sau khi được hoàn chỉnh các mục từ và định dạng theo cấu trúc từ điển và câu trúc luật mã hóa.

Ví dụ câu đầu vào tiếng Việt "Long là bác sĩ của cô ấy", để công cụ EnCoVie chuyển đổi sang biểu thức UNL thì chúng tôi cần cung cấp:

- Các mục từ trong từ điển tiếng Việt – UNL

```
[cô ấy] {} "she(icl>person)" (P,NP,sg3) <vie,0,0>;
[bác sĩ] {} "doctor(icl>medical_practitioner)"(P,NP,sg3)<vie,0,0>;
[của] {} "" (E) <vie,0,0>;
[Long] {} (N,NP) <vie,0,0>;
```

- Các luật trong tập luật chuyển đổi ngữ pháp

```
-{"là":null:null}{N:+.@present:null};
>{N,NP:null:aoj}{N,@present:null:null};
- {E:null:null}{N:+@pos:null};
<{N:null:null}{P,@pos:null:pos};
```

Giao diện của hệ thống EnCoVie gồm có bốn cửa sổ: cửa sổ thứ nhất chứa các tập luật mã hóa, cửa sổ thứ hai chứa bộ từ điển tiếng Việt – UNL, cửa sổ thứ ba là vùng nhập câu tiếng Việt, cửa sổ thứ tư là trình bày các bước chuyển đổi để theo dõi quá trình mã hóa và kết quả đầu ra.

Hình 4. Công cụ chuyển đổi câu tiếng Việt sang biểu thức UNL

3. Thử nghiệm

Chúng tôi tiến hành thử nghiệm trên công cụ EnCoVie với khoảng 100 câu đơn tiếng Việt được dịch sang UNL. Để đánh giá chất lượng đầu ra của biểu thức UNL, chúng tôi tiếp tục thực hiện thử nghiệm như sau:

- Dữ liệu thử nghiệm: chúng tôi xây dựng hai bộ dữ liệu gồm 100 câu tiếng Việt làm dữ liệu đầu vào và 100 câu tiếng Anh làm dữ liệu đích (câu mẫu so sánh)
- Lần thử nghiệm thứ nhất: chúng tôi sử dụng công cụ EnCoVie dịch câu tiếng Việt sang UNL, kết quả đầu ra tiếp tục được dịch sang tiếng Anh bởi một công cụ được đặt tại địa chỉ http://www.unl.ru/deco.html.
- Lần thử nghiệm thứ hai: chúng tôi sử dụng công cụ Google translate (https://translate.google.com.vn/) để dịch trực tiếp từ tiếng Việt sang tiếng Anh.

Kết quả của hai lần thử nghiệm được chúng tôi mô tả qua bảng thống kê sau:

Bảng 1. Tỷ lệ thay đổi của hai phương pháp dịch

Phương pháp dịch	Tỷ lệ thay đổi của câu	
	Giống câu mẫu	Không giống câu mẫu
Tiếng Việt →UNL→tiếng Anh	80%	20%
Tiếng Việt → tiếng Anh	66.67%	33.33%

B. Đánh giá

Chúng tôi đã nghiên cứu cấu trúc các mô hình tiêu biểu câu đơn tiếng Việt, các mối quan hệ, thuộc tính của các từ loại trong tiếng Việt và xác định sự tương đương giữa các từ loại, thuộc tính trong câu tiếng Việt với các thuộc tính của UW, các quan hệ giữa UWs trong biểu thức UNL. Chúng tôi đã chọn ra một mô hình để phân tích và xây dựng tập dữ liệu câu tiếng Việt. Từ đó, chúng tôi tiến hành xây dựng tập luật để mã hóa các dạng câu này sang biểu thức UNL.

Chúng tôi xây dựng module dùng để tách từ, gán nhãn từ loại cho câu đầu vào tiếng Việt và tích hợp vào công cụ EnCoVie. Chúng tôi thử nghiệm khoảng 100 câu trên công cụ EnCoVie để chuyển đổi tiếng Việt sang UNL và từ UNL sang tiếng Anh. Cùng trên tập dữ liệu đó, chúng tôi sử dụng công cụ Google translate dịch trực tiếp từ tiếng Việt sang tiếng Anh. Qua bảng 1, chúng tôi thấy dịch qua ngôn ngữ trung gian (UNL) chất lượng bản dịch cuối cùng tốt hơn so với dịch trực tiếp (bằng phương pháp thống kê). Ví dụ với câu đầu vào tiếng Việt "nhà tôi có khách" và câu đích tiếng Anh "my house has guest", nếu dịch qua UNL thì câu đích là "My house has a guest" và dịch trực tiếp là "I have guest houses". Tuy nhiên, tập dữ liệu thử chúng tôi xây dựng còn nhỏ chỉ giới hạn trên một số cấu trúc câu đơn tiếng Việt nên đánh giá có tính thử nghiệm.

Với con số 20% không giống mẫu, chúng tôi thống kê là do hai nguyên nhân chính:

- Module tách từ và gán nhãn từ loại không chính xác cho một số từ loại trên một số câu tiếng Việt nên dẫn đến công cụ EnCoVie không tìm ra luật hoặc tìm sai luật. Ví dụ câu tiếng Việt "nhà này năm tầng" với nhãn từ loại ("nhà" là danh từ, "này" là đại từ, "năm" là số từ và "tầng" là danh từ) nhưng công cụ lại gán nhãn từ "này" lại là số từ.
- Quá trình tìm luật của công cụ EnCoVie là tìm theo thứ tự từ trên xuống trong tập luật nên có lúc nhận dạng sai luật. Nguyên nhân là do: dữ liệu trong luật thiếu thông tin, thiếu thông tin các thuộc tính trong bộ từ điển tiếng Việt – UNL và chúng tôi chưa xác định mức ưu tiên cho từng luật.

V. KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày thuật toán xác định mối quan hệ giữa các UWs, xác định các thuộc tính cho các UWs khi chuyển đổi một câu đơn tiếng Việt sang biểu thức UNL. Chúng tôi đã xây dựng một công cụ mã hóa câu tiếng Việt sang biểu thức UNL gọi là EnCoVie.

Trong quá trình nghiên cứu câu đơn tiếng Việt, chúng tôi thấy rằng không phải ở mô hình nào cũng có sự khớp nhau giữa cách phân theo cấu trúc và phân theo ngữ nghĩa vì vậy việc quy chúng thành những vị trí ổn định của mô hình không phải lúc nào cũng rành mạch do đó sẽ có rất nhiều mô hình biến thể và chúng ta cũng không thể lấy từ loại để định vị cho mô hình.

So với các công cụ mã hóa cho tiếng Việt như IAN, EnCo [11] thì công cụ EnCoVie được xây dựng có cấu trúc luật mã hóa và giao diện thân thiện với người dùng. Cửa sổ nhập luật, từ điển, câu đầu vào và kết quả đầu ra chỉ thực hiện trên một trang giao diện. Kết quả chuyển đổi từ câu tiếng Việt sang UNL của công cụ EnCoVie được chúng tôi so sánh với các câu thực hiện thủ công thì tương đương khoảng gần 80% tổng số câu giống nhau. Chúng tôi sẽ tiếp tục hoàn thiện công cụ EnCoVie về các nội dung: bổ sung thêm thông tin cho luật mã hóa câu đơn, thông tin thuộc tính của các mục từ trong bộ từ điển và xác định mức độ ưu tiên thực hiện luật. Kết quả trên sẽ là cơ sở để chúng tôi nghiên cứu các dạng mô hình khác của câu đơn, câu ghép tiếng Việt và thời, thì trong tiếng Việt.

TÀI LIỆU THAM KHẢO

[1] P. T. L. Thuyen, V. T. Hung, "Results comparison od machine translation by direct translation and by through intermediate language", International Journal of Advance Research in Computer Science and Management Studies, Volume 3, Issue 4, April 2015.

[2] M. Zhang, X. Duan, V. Pervouchine, H. Li, "Machine Transliteration: Leveraging on Third Languages", Coling 2010: Poster Volume, pages 1444–1452,Beijing, August 2010.

[3] UNL centre, "Enconverter Specifications", Version 3.3, http://www.undl.org, 2002.

[4] P. Kumar and R. K. Sharma, "Generation of unl attributes and resolving relations for punjabi enconverter", Malaysian Journal of Computer Science, Vol. 24(1), pages 34-46, 2011.

[5] M. Jain M, O. P. Damani, "English to UNL (Interlingua) Enconversion", Proc. 2nd Conference on Language and Technology, Lahore, Pakistan, papes 1-8, 2009.

[6] J. Balaji, T. V. Geetha, Ranjani and M. Karky, "Morpho-Semantic Features for Rule-based Tamil Enconversion", International Journal of Computer Applications, Volume 26– No.6, pp. 0975-8887, 2011.

[7] Md. N. Y. Ali, A. M. Nurannabi, M. A. Ali, J. K. Das, G. F. Ahmed, "Conversion of Bangla Sentence for Universal Networking Language", Proceedings of 13th International Conference on Computer and Information Technology (ICCIT 2010), pages 108–113,2010.

[8] H. T. Phiến, "Ngữ pháp tiếng Việt", Nhà xuất bản Đại học và Trung học chuyên nghiệp, 1980.

[9] Vo-Trung H., Fafiotte G., "UVDict – a machine translation dictionary for Vietnamese language in UNL system", Proceeding CISIS 2011, Korean Bible University (KBU), Seoul, Korea, Pages 1020-1028, 2011.

[10] Phan L. T., Vo-Trung H., "Expand data on UNL – Vietnamese dictionnary of UNL Explorer", Journal of Science and Technology, University of Danang, No 56, 2014.

[11] P. T. L. Thuyen, V. T. Hung, "Automatic translation for Vietnamese based on UNL language", International Conference on Electronics Information and Communacation (ICEIC), page 628- 632, 2016.

CONVERSION OF A VIETNAMESE SIMPLE SENTENCE INTO UNL EXPRESSION

Phan Thi Le Thuyen, Vo Trung Hung

ABSTRACT— *UNL (Universal Networking Language) is an artificial language and can express any content of natural language in the manner of it. UNL is the basis for the development of multilingual automatic translation software through a pivot language (in this case is the UNL language). UNL opens the possibility that the user can access information on the Internet without language barriers. UNL has been studied and applied for 48 languages. UNL system consists of two main components as encoding (EnCoverter) and decoding (DeConverter). The encoding is the process of converting a text from the source language (eg English, Vietnamese, ...) into a text written in UNL, and the decoding is the reverse transition (from UNL language to the target language). Currently, UNL systems for Vietnamese is undeveloped. In this paper, we present the results of research on the method of encoding a Vietnamese simple sentence into corresponding UNL expression. To make the transition, we have to build Dictionary Vietnamese - UNL, the laws (rules of grammar), and software to convert. Especially, we have proposed an algorithm using the decoding laws to create the corresponding property of the UNL and resolve other relationships when converting. Achievements that we have built tools EnCoVie that perform decoding functions for some cases of Vietnamese simple sentences.*

Keywords— *Automatic translation, UNL, Multilingual translation, natural language processing, coding.*