Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00089

PHÂN TÍCH CÚ PHÁP CÂU ĐƠN TIẾNG VIỆT ĐỂ XÂY DỰNG CÔNG CỤ DECONVERTER

Phan Thị Lệ Thuyền, Võ Trung Hùng
Đại học Đà Nẵng
thuyenptl@gmail.com, vthung@dut.udn.vn

TÓM TẮT: *Ngôn ngữ UNL (Universal Networking Language) đã được nhiều nhà nghiên cứu sử dụng như một ngôn ngữ trung gian trong dịch tự động. Hệ thống dịch tự động dựa vào UNL bao gồm hai thành phần chính là EnConverter và DeConverter. DeConverter sử dụng để tạo ra ngôn ngữ tự nhiên từ biểu thức UNL dựa vào bộ từ điển UNL - ngôn ngữ tự nhiên và tập luật giải mã. Trong bài báo này, chúng tôi trình bày giải pháp xây dựng công cụ chuyển đổi một biểu thức UNL có một hay nhiều nút con sang câu tiếng Việt tương ứng gọi là DeCovie. Chúng tôi tiến hành phân tích ngữ pháp giữa biểu thức UNL có một hoặc nhiều nút con và câu đơn trong tiếng Việt và đề xuất loại luật để chuyển đổi từ biểu thức UNL sang câu đơn tiếng Việt. Chúng tôi sử dụng 200 biểu thức UNL để chuyển đổi sang tiếng Việt, kết quả dịch rất tốt so với các công cụ Deconverter khác.*

Từ khóa: *Dịch máy, Ngôn ngữ mạng dùng chung, dịch liên ngôn ngữ, xử lý ngôn ngữ tự nhiên, giải mã.*

I. GIỚI THIỆU

Dịch tự động là quá trình dịch từ ngôn ngữ này (*ngôn ngữ nguồn*) sang một hay nhiều ngôn ngữ khác (*ngôn ngữ đích*) một cách tự động mà không có sự can thiệp của con người [14]. Có nhiều phương pháp trong dịch tự động, trong đó phương pháp dịch qua ngôn ngữ trung gian có ưu điểm là giảm số lượng cặp dịch từ n*(n-1)/2 xuống 2*n nếu hệ thống có n ngôn ngữ tham gia và giải quyết cho các cặp dịch thiếu về tài nguyên ngôn ngữ [1]. Tuy nhiên, để tìm được một ngôn ngữ làm trung gian có thể biểu diễn tất cả thông tin của mọi ngôn ngữ tự nhiên mà không gây nhập nhằng là vô cùng khó khăn [14].

Năm 1996, TS. Hiroshi Uchida đề xuất một ngôn ngữ nhân tạo gọi là UNL. Ý tưởng cơ bản của UNL là định nghĩa ra một ngôn ngữ như là một ngôn ngữ trục (*pivot language*) có khả năng biểu diễn tất cả các ngôn ngữ tự nhiên mà không bị nhập nhằng về ngữ nghĩa. Mục đích chính của UNL 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 dịch tự động dựa trên UNL đã cung cấp hai công cụ EnCoverter và DeConverter. Công cụ DeCoverter thực hiện chức năng chuyển một văn bản được viết trong UNL sang văn bản được viết trong ngôn ngữ tự nhiên. Có nhiều công cụ DeConverter được xây dựng như: EUGENE và DeCo cung cấp nền tảng UNL để các nhà phát triển ngôn ngữ mở rộng cho ngôn ngữ của họ [3][13], giải mã từ biểu thức UNL sang tiếng Punjabi [4], giải mã từ biểu thức UNL sang tiếng Bangla [5],... Đối với tiếng Việt, hiện có nhiều nghiên cứu về khả năng ứng dụng UNL trong dịch tự động như: xây dựng từ điển tiếng Việt - UNL và UNL - tiếng Việt [7] [8] [15], sử dụng các công cụ có sẵn để dịch câu tiếng Việt sang UNL và ngược lại [9], xây dựng công cụ EnConverter để dịch câu đơn và câu ghép tiếng Việt sang UNL [10-11].

Cộng đồng các nhà nghiên cứu dịch tự động dựa trên UNL đã cung cấp hai công cụ EUGENE và DeCo cho tất cả các ngôn ngữ tự nhiên có thể tham gia vào hệ thống. Tuy nhiên trên thực tế, các công cụ này thích hợp cho việc thử nghiệm hơn là mở rộng hệ thống vì chúng được định nghĩa rất nhiều loại luật sử dụng để chuyển đổi (*trên 18 loại luật*) và không có tính kế thừa các nghiên cứu của ngôn ngữ mà chúng ta muốn mở rộng. Giải pháp các nhà phát triển chọn là xây dựng công cụ mới cho ngôn ngữ của họ. DeConverter đã được nghiên cứu và ứng dụng cho hơn 54 ngôn ngữ khác nhau, tuy nhiên hiện nay chưa có công cụ nào được xây dựng riêng cho tiếng Việt.

Trong bài báo này, chúng tôi trình bày kết quả nghiên cứu về phương pháp chuyển đổi biểu thức UNL có một hay nhiều nút con sang câu tiếng Việt tương ứng. Để xây dựng công cụ DeConverter riêng cho tiếng Việt, chúng tôi phân tích cú pháp biểu thức UNL và phân tích ngữ pháp câu đơn tiếng Việt để đề xuất các loại luật cho công cụ và xây dựng luật chuyển đổi cho các quan hệ được định nghĩa trong UNL. Chúng tôi đề xuất giải pháp phân tích biểu thức UNL thành các quan hệ, UW và tạo liên kết với mục từ bộ từ điển UNL - tiếng Việt [8].

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 đề xuất của chúng tôi để chuyển đổi một biểu thức UNL có một hoặc nhiều nút con sang câu tiếng Việt ; 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. Ngữ pháp tiếng Việt

Tiếng Việt thuộc loại hình ngôn ngữ đơn lập, vì vậy từ trong tiếng Việt sẽ không có biến đổi hình thái như các ngôn ngữ khác (*ví dụ tiếng Anh*). Do đó các phương thức ngữ pháp tiếng Việt chủ yếu dựa vào trật tự từ, hư từ và ngữ điệu. Phương thức trật tự từ là sự sắp xếp các từ theo một thứ tự nào đó để biểu thị quan hệ ngữ pháp của câu (*"sau*

nhà" và "nhà sau"). Phương thức hư từ là sử dụng các từ để thể hiện nhưng các hư từ này không có khả năng tạo thành phần câu khi hoạt động độc lập (*"mẹ của con" và "mẹ và con"*). Phương pháp ngữ điệu thường sử dụng các dấu câu để biểu hiện như muốn đưa ra một nội dung thông báo (*ai gõ cửa thế?*). Do đó, việc phân tích ngữ pháp cho câu trong tiếng Việt là dựa hoàn toàn vào trật tự tự của các từ trong câu và có nhiều bài toán tách từ, gán nhãn từ loại và phân tích ngữ pháp tiếng Việt được nghiên cứu có kết quả chính xác cao [12,13].

Trong tiếng Việt, kết cấu chủ - vị (*C-V*) là đơn vị cú pháp nhỏ nhất tức là kết cấu có hai vế được đặt theo quan hệ cú pháp cơ bản là quan hệ chủ ngữ và vị ngữ [6]. Câu đơn trong tiếng Việt là loại câu cơ sở, phần lớn câu đơn tiếng Việt ứng với một kết cấu C-V và mang thông tin ngữ nghĩa tự thân. Sự khác nhau về câu đơn tiếng Việt không phải là sự khác nhau về độ ngắn dài mà là khác nhau tính chất các quan hệ trong cấu trúc.

B. Biểu thức UNL

UNL là một ngôn ngữ nhân tạo được phát triển dựa vào phương pháp trung gian cho dịch máy. Nó cho phép người sử dụng có thể biểu diễn tất cả các tri thức của ngôn ngữ tự nhiên mà không nhập nhằng. UNL bao gồm các thành phần của ngôn ngữ tự nhiên: từ vựng (*Universal Word -UW*), quan hệ *(relation*), thuộc tính (*attributes*) và cơ sở tri thức ngôn ngữ (*UNL Knowledge Base -UNLKB*). 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 và thể hiện quan điểm của người nói được diễn đạt, tri thức ngôn ngữ dùng để định nghĩa các quan hệ có thể giữa các khái niệm trong ngôn ngữ [3].

Một biểu thức UNL là một tập hợp các quan hệ nhị phân, mỗi quan hệ nhị phân bao gồm một quan hệ để liên kết hai từ vựng. Một biểu thức UNL là một danh sách các mối quan hệ nhị phân được mô tả chung như sau:

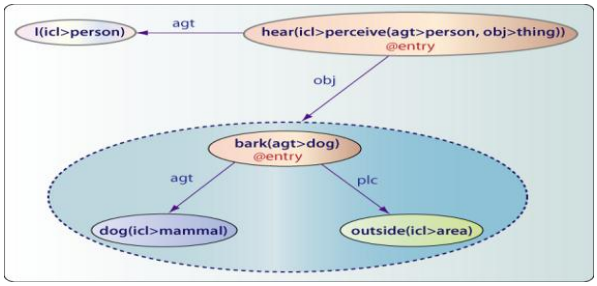
```
<relation> [:<scope-ID>] ( <from-node>, <to-node> )
```

Trong đó: <relation> là một trong các nhãn quan hệ và <from-node>, <to-node> là hai từ vựng có quan hệ nhau qua <relation>, <scope-ID> là một nút phạm vi được định nghĩa.

Ví dụ một câu đầu vào tiếng Anh "*I can hear a dog barking outside*" được chuyển sang UNL tương đương 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 đó: "agt", "obj" và "plc" là các quan hệ; "hear(icl>perceive(agt>person,obj>thing)", "I(icl>person", "bark(agt>dog)", "dog(icl>mammal)"; "outside(icl>area)" là các từ vựng; "@entry", "@ability", "@topic", "@progress" và "@indef" là các thuộc tính; ":01" được định nghĩa như là nút kết hợp (scope)



Hình 1. Đồ thị của biểu thức trong UNL

C. Cấu trúc từ điển UNL - tiếng Việt

Một mục từ của từ điển được định dạng trong từ điển UNL - tiếng Việt như sau[8]:

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

Trong đó: "*HW*" là từ đầu mục từ của ngôn ngữ; "*UW*" là từ vựng; "*ATTR*" là thuộc tính ngữ pháp; "*FLG*" là cờ ngôn ngữ; "*FRE*" là tần số xuất hiện và "*PRI*" là mức ưu tiên thực hiện luật.

D. Một số quan hệ được định nghĩa trong UNL

1. Quan hệ "*aoj*" được định nghĩa một sự việc mà đang ở trạng thái hoặc thuộc tính nào đó. Ví dụ ta có quan hệ "*aoj*" gắn kết giữa hai từ vựng "*nice*" và "*ski(agt>person)*" như sau: aoj(nice,ski(agt>person). Phá vỡ mối quan hệ trên để chuyển thành câu tiếng Anh tương ứng là "*Skiiing is nice*". Cấu trúc định nghĩa trường hợp này:

```
aoj(UW1=thing,UW2=thing)
```

2. Quan hệ "*agt*" được định nghĩa sự việc khởi đầu cho một hành động. Ví dụ ta có quan hệ "*agt*" gắn kết giữa hai từ vựng '*translate(agt>thing,gol>language,obj>information,src>language*)" và "*computer(icl>machine))*" như sau: agt(translate(agt>thing,gol>language,obj>information,src>language),computer(icl>machine)). Phá vỡ mối quan hệ trên để chuyển thành câu tiếng Anh tương ứng là "*computer translates*...". Cấu trúc định nghĩa trường hợp này:

```
agt(UW1=do,UW2=thing)
```

3. Quan hệ "*obj*" được định nghĩa một việc trung tâm bị ảnh hưởng trực tiếp bởi một sự kiện hoặc trạng thái. Ví dụ ta có quan hệ "*obj*" gắn kết giữa hai từ vựng "*move(gol>place,obj>thing,src>place)"* và "*table(icl> furniture)*" như sau: obj(move(gol>place,obj>thing,src>place),table(icl> furniture)). Phá vỡ mối quan hệ trên để chuyển thành câu tiếng Anh tương ứng là "*the table moved*". Cấu trúc định nghĩa trường hợp này:

```
obj(UW1=do,UW2=thing)
```

4. Quan hệ "*plc*": định nghĩa nơi mà sự kiện xảy ra hoặc một trạng thái là đúng hoặc một sự việc. Ví dụ ta có quan hệ "*plc*" gắn kết giữa hai từ vựng "*cook(icl>do)"*và*"kitchen(pof>building)"* như sau: plc(cook(icl>do),kitchen (pof>building)). Phá vỡ mối quan hệ trên để chuyển thành câu tiếng Anh tương ứng là *"... cook ... in the kitchen*". Cấu trúc định nghĩa trường hợp này:

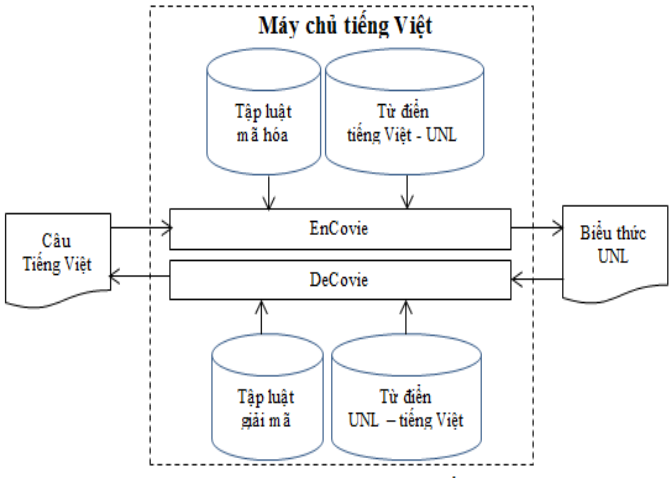
```
obj(UW1=thing,UW2=thing)
```

III. GIẢI PHÁP ĐỀ XUẤT

A. Hệ thống dịch UNL cho tiếng Việt

Hệ thống dịch tự động đa ngữ UNL bao gồm nhiều máy chủ ngôn ngữ (Language Servers - LSs) 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. Máy chủ ngôn ngữ sẽ đảm nhận 2 chức năng đó là dịch một văn bản từ ngôn ngữ này sang ngôn ngữ UNL (*gọi là EnConverter*) và dịch ngược lại (*gọi là DeConverter*).

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 2. Hệ thống UNL cho tiếng Việt

B. Định dạng luật giải mã

Kết quả phân tích cấu trúc ngữ pháp câu đơn trong tiếng Việt và quan hệ trong UNL [3, 9], trong bài báo này chúng tôi đề xuất 2 loại luật giải mã sử dụng để chuyển đổi từ biểu thức UNL sang câu đơn tiếng Việt như sau:

1. *Luật chèn phải*. Luật này chèn một nút mới vào bên phải cửa sổ LGW trên Node-list, quá trình chèn có thể thay đổi thuộc tính của nút.

```
:{<COND1>:<ACTION1>:<RELATION1>}"<COND2>:<ACTION2>:<RELATION2>";
```

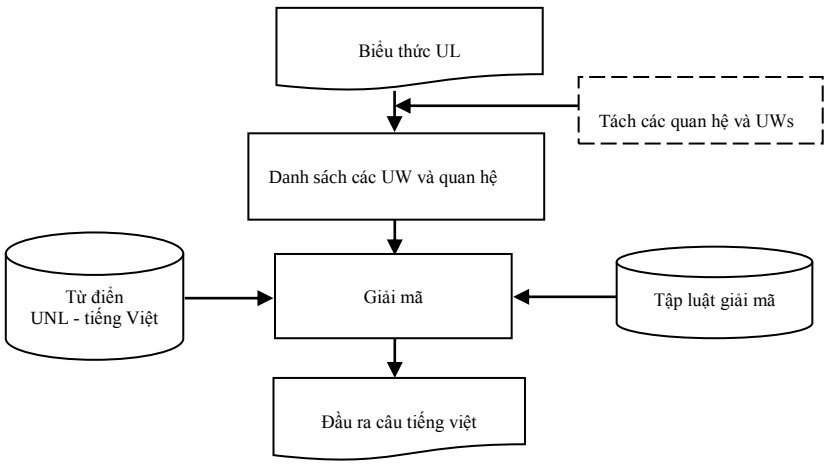
Trong đó, <COND1> và <COND2> chỉ điều kiện 1 và 2, chứa các thuộc tính từ vựng và ngữ nghĩa của cửa sổ phân tích trái và phải. <ACTION1> và <ACTION2> chỉ hành động được thực hiện nếu điều kiện tương ứng đúng. <REL1> và <REL2> chỉ ra mối quan hệ có thể có giữa hai cửa sổ phân tích.

2. *Luật chèn trái*. Luật này chèn một nút mới vào bên trái cửa sổ LGW trên Node-list, quá trình chèn có thể thay đổi thuộc tính của nút.

```
:"<COND1>:<ACTION1>:<RELATION1>"{<COND2>:<ACTION2>:<RELATION2>};
```

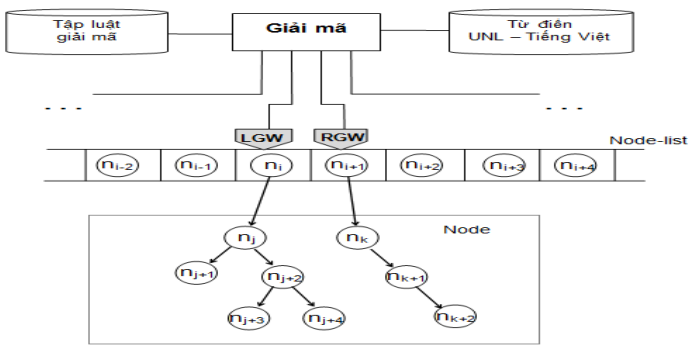
C. Công cụ DeCovie

Hoạt động của DeCovie được mô tả như sau: Với biểu thức UNL đầu vào là tập hợp các mối quan hệ nhị phân, DeCovie sẽ phá vỡ các mối quan hệ nhị phân đó bằng cách thực hiện việc tách các quan hệ và UW. Các UW được lưu trên danh sách Node (n_1,n_2,...n_n), các mối quan hệ được lưu trên danh sách Rel (r_1,r_2,...r_n) và liên kết với Node. Quá trình giải mã được thực hiện dựa vào bộ từ điển UNL - tiếng Việt và tập luật giải mã.



Hình 3. Sơ đồ chuyển đổi biểu thức UNL sang câu tiếng Việt

Danh sách Node-list có hai cửa sổ hoạt động dịch chuyển và chèn nút, cửa sổ bên trái gọi là LGW (*Left Generation Window*) và cửa sổ bên phải gọi là RGW (*Left Generation Window*). LGW và RGW kiểm tra hai nút nếu thỏa mãn luật nào đó thì thực hiện. Khi quá trình kết thúc, các nút trên Node -list chứa câu tiếng Việt đầu ra.



Hình 4. Hai cửa sổ hoạt động trên Node-list

1. *Giải pháp phân tích cú pháp biểu thức UNL*

Mỗi nút đồ thị của UNL đại diện cho khái niệm của UW và mỗi cạnh là đại diện cho một quan hệ nhị phân giữa hai nút. Các cạnh của đồ thị sẽ có hướng đi từ nút cha cho đến nút con, nghĩa là trong một mối quan hệ nhị phân của UNL thì UW2 được xem là phụ thuộc vào UW1. Nút gốc của đồ thị được gọi là nút vào (*entry node*) và được xác định bởi thuộc tính @entry. Giải pháp phân tích cú pháp biểu thức UNL được trình bày như sau:

Bước 1. Khởi tạo hai danh sách node và rel.

Bước 2. Đọc dòng văn bản thứ i

Bước 3. Tách các UWs ra khỏi các quan hệ. Kiểm tra các UWs trong danh sách node.

+ Nếu chưa tồn tại, $node_i$=UW_1 và $node_{i+1}$ = UW_2. Đánh dấu rel_i= tên quan hệ (UW_1,UW_2)

+ Nếu đã tồn tại, Đánh dấu rel_i= tên quan hệ (UW_1,UW_2)

Bước 4. Quay lại bước 2

Thuật toán C.1. Phân tích cú pháp biểu thức UNL

Input: Biểu thức UNL

Output: Các UWs và quan hệ

```
1. {
2.    Tạo node list; rel list;
3.    While (UNLexp not null)
4.  {
5.   Read (UNLexp);
6.   i=1;
7.   rel_i = quan hệ (UW_1;UW_2);
8.   If (UW1 not in node)
9.     {
10.      thêm UW_1 vào node_i;
11.      ++i;
12.    }
13.  If (UW_2 not in node)
14.     thêm UW_1 vào node_i;
15.  ++i;
16.  }
return(output);}
```

2. Chuyển đổi biểu thức UNL sang tiếng Việt

Bước 1. Phân tích cú pháp biểu thức UNL (*được trình bày ở GT C.1*)

Bước 2. Gắn liên kết giữa các nút và các mục từ trong từ điển tiếng Việt - UNL (*nếu không tìm thấy mục từ trong từ điển thì sử dụng chính từ gốc và gắn thuộc tính Unattri*).

Bước 3. Tìm UW có gắn thuộc tính @entry (*gọi là UW đầu vào của biểu thức*).

Bước 4. Khởi tạo danh sách chứa câu đầu ra gọi là node-list. Trạng thái ban đầu của node-list bao gồm 3 nút: Nút bắt đầu câu (<<), nút đầu vào chứa UW có thuộc tính @entry và nút kết thúc câu (>>).

Bước 5. Khi bắt đầu quá trình giải mã, nút đầu câu sẽ LGW và nút kế tiếp sẽ là RGW. LGW và RGW hoạt động trên các nút từ trái sang phải của danh sách node.

+ Nếu tìm thấy luật, chèn nút chứa kết quả đầu ra (*của ngôn ngữ tự nhiên*) vào node-list và dựa vào các thuộc tính của UW để xác định ngữ pháp của kết quả đầu ra (*ví dụ xác định yếu tố thời gian cho động từ*). Xóa bỏ các mối quan hệ giữa hai UW của các nút trên node. LGW và RGW di chuyển sang trái một nút.

+ Nếu không tìm thấy luật thì chuyển sang bước 7.

Bước 6. Nếu như LGW chạm vào nút cuối câu (>>) thì hệ thống ngừng xử lý và xóa thuộc tính .@entry của nút gốc. Nó có nghĩa quá trình giải mã một biểu thức UNL sang tiếng Việt được xử lý thành công bởi hệ thống.

Bước 7. Dịch chuyển LGW và RGW di chuyển sang phải một nút (*nghĩa là LGW sẽ trở thành RGW và nút kế tiếp sẽ trở thành RGW*). Quay lại bước 5.

Thuật toán C.2. Chuyển đổi biểu thức UNL sang tiếng Việt.

Input: node{n_1,n_2,…,n_n} và rel{r_1,r_2,...,r_n};

Output: Câu tiếng Việt

```
1.{
2. link(node,dict);
3. Tạo node-list chứa 3 nút ({<<}, {nroot}{>>});
4. LGW =node-list1;
5. RGW=node-list2;
6. While (LGW = "<<") do
7.   If ({tìm luật})
8.    {
9.      If (luật chèn trái)
10.      {
11.     Chèn nút vào bên trái RGW trên node-list;
12.     Thêm thuộc tính cho nút;
13.     Xóa mối quan hệ trong rel;
14.     LGW → trái;
15.     RGW → trái;
16.     Xóa nút trong node;
17.     }
18.    Else
19.     {
20.     Chèn nút mới vào bên phải RGW;
21.     Thêm thuộc tính cho nút;
22.     Xóa quan hệ trong rel;
23.     LGW → trái;
24.     RGW → trái;
25.     Xóa nút trong node;
26.     }
27.  Else
28.   {
29.    RGW → phải;
30.    LGW→ phải;
31.   }
32. Xóa thuộc tính trong node-list;
Output (node-list);}
```

3. Giải pháp trường hợp biểu thức UNL có một hoặc nhiều nút con

Bước 1. Bắt đầu từ nút gốc "*@entry*" của biểu thức. Xác định các nút con và các mối quan hệ với các nút.

Bước 2. Từ nút gốc và xem nút này là nút cha lớn nhất, hệ thống đi thăm một nút con. Kiểm tra nút này có nút con khác không. Nếu không thì chuyển sang bước 4. Nếu có thì chuyển sang bước 3.

Bước 3. Chèn nút vào node-list và xem nút này là nút cha của các quan hệ con của nó. Quay lại bước 2.

Bước 4. Hệ thống thực hiện lệnh chèn nút vào node-list, xóa nút vừa chèn. Hệ thống sẽ trở về nút cha và tiếp tục bước 2.

Thuật toán C.3. Chuyển đổi biểu thức UNL có nhiều nút con sang tiếng Việt.

Input: Node{$n_1,n_2,\ldots,n_n$} và rel{$r_1,r_2,...,r_n$};

Output: Câu đơn tiếng Việt

```
1.{
2. node-list =n_root;
3. While (node not null)
4.  {
5.    n_i=n_root;
6.    While (n_i.link not null) do
7.     {
8.       Visit(n_i.link);
9.       If ({n_i.link not null})
10.        {
11.        Chèn n_i vào node-list;
12.          Xóa n_i;
13.         n_i=n_i.link;
14.        }
15.      Else
16.        {
17.          Chèn n_i vào node-list;
18.            Xóa n_i;
19.        }
20.     }
21.  }
Output (node-list);}
```

D. Xây dựng luật giải mã

Trong bài báo này, chúng tôi phân tích trên hai trường hợp biểu thức UNL chuyển sang câu tiếng Việt như sau:

- *Trường hợp biểu thức có một nút con với quan hệ "aoj"*

Quan hệ ngữ nghĩa "*aoj*" trường hợp này với hai từ vựng có thuộc tính lần lượt là "*n, nt*" và "*p, pp*".

Danh từ (*n*) có khả năng làm thành tố chính trong cụm danh từ, danh từ thường làm chủ ngữ trong câu . Danh từ đơn thể (*nt*) thường chỉ các quan hệ thân thuộc, chức vụ, nghề nghiệp. Đại từ (*p*) là lớp từ dùng để thay thế và chỉ trỏ, ý nghĩa thay thế ở đây là thay thế cái đã được gọi tên, cái đã được nói đến và được biết trước đó. Đại từ có thể đảm nhận được các chức năng cú pháp của thực từ thay thế. Đại từ nhân xưng (*pp*) dùng để thay cho người nói được chia thành các ngôi rõ rệt: ngôi thứ nhất (*tôi, tao, ta, tớ*), ngôi thứ hai chỉ người đồng thoại (*mày, cậu, mi*), ngôi thứ ba chỉ người hoặc sự vật được nói đến (*nó, hắn, y, ả, ta, chúng ta, chúng tôi*), ngôi thứ hai số nhiều (*chúng mày, chúng bay*), ngôi thứ ba số nhiều (*họ, chúng, chúng nó*).

Luật phá vỡ mối quan hệ giữa hai UW và chèn, sắp xếp trật tự các từ trong câu đích như sau:

```
:"n,nt:null:aoj"{p,pp:null:null};
```

Ví dụ với trường hợp tổng quát trên, ta có biểu thức UNL sau:

```
{unl}
aoj(teacher(icl>educator>thing).@entry.@present.@affirmative,I(icl>person))
{/unl}
```

Trong cấu trúc này có thuộc tính "*@affirmative*" - nó mang tính chất khẳng định và gắn với hệ từ "*là*" trong tiếng Việt. Thuộc tính "*@affirmative*" gắn UW có thuộc tính "*n, nt*" tạo một cụm từ trong tiếng Việt và hệ từ "*là*" đứng trước để biểu thị ý nhấn mạnh sắc thái khẳng định. Luật tạo thành ngữ pháp tiếng Việt tương đương như sau:

```
:"[là]:+C:null"{n,nt.@affirmative:-@present,@affirmative :null};
```

Để tạo thành một câu đích hoàn chỉnh, quá trình chuyển đổi còn sử dụng các luật để tạo khoảng cách trắng giữa các từ trong câu như sau:

```
:{p,pp,^blk:+blk:null}"[],blk:null:null";
:{C,^blk:+blk:null}"[],blk:null:null";
```

→ Câu tiếng Việt đầu ra tương đương biểu thức UNL "*tôi là giáo viên*".

- Trường hợp biểu thức chứa ba nút con với quan hệ "agt", "obj" và "plc"

Quan hệ ngữ nghĩa "*agt*" với hai từ vựng có thuộc tính lần lượt là "*v, vt*" và "*p, pp*".

Khi quan hệ "*agt*" với UW1 có thuộc tính "*v, vt*" và UW2 có thuộc tính "*p, pp*". Ngoại động từ (*vt*) là những động từ chỉ hoạt động có bắc cầu sang những sự vật ở ngoài nó. Khi những động từ này làm chính tố (*vị ngữ chính*) trong câu đòi hỏi phải có bổ ngữ đối tượng trực tiếp, bổ ngữ đối tượng gián tiếp, hoặc bổ ngữ tự do thì nghĩa của câu mới được hiểu đầy đủ. Đại từ (*p*) có thể đảm nhận được các chức năng cú pháp của thực từ thay thế. Luật phá vỡ mối quan hệ giữa hai UW và chèn, sắp xếp trật từ các từ trong câu đích như sau:

```
:"v,vt:null:agt"{p,pp:null:null};
```

Khi quan hệ "obj" với UW1 có thuộc tính "v, vt" và UW2 có thuộc tính "n, na" với cấu trúc. Luật phá vỡ mối quan hệ giữa hai UW và chèn, sắp xếp trật tự các từ trong câu đích như sau:

```
:{v,vt:null:null}"n,na:null:obj";
```

Khi quan hệ "plc" với UW1 có thuộc tính "n, na" và UW2 có thuộc tính "n, ng". Về khả năng kết hợp thì danh từ nó có khả năng làm chính tố trong cụm danh từ. Luật phá vỡ mối quan hệ giữa hai UW và chèn, sắp xếp trật tự các từ trong câu đích như sau:

```
:{n,na:null:null}"n,ng:+@plc:plc";
```

Ví dụ ta có biểu thức UNL sau:

```
{unl}
agt(send(icl>direct>do,plt>uw,plf>thing,agt>volitional_
    thing,obj>thing,rec>thing).@entry.@present,i(icl> person))
obj(send(icl>direct>do,plt>uw,plf>thing,agt>volitional_thing,obj>thing,rec>thing).@entry
    .@present,letter(icl> text>thing))
plc(letter(icl>text>thing),post(icl>upright>thing))
{/unl}
```

Nếu một từ có thuộc tính "*@plc*" là dấu hiệu nhận biết đây là một từ chỉ địa điểm, trong tiếng Việt sẽ thêm giới từ để biểu thị điều sắp nêu ra là nơi, chỗ, khoảng thời gian sự vật hay sự việc được nói đến tồn tại hay diễn ra. Luật thêm giới từ tạo quan hệ ngữ pháp tiếng Việt như sau:

```
:"[ở]:+e:null:null"{n,ng,@plc:-@plc :null};
```

Để tạo thành một câu đích hoàn chỉnh, quá trình chuyển đổi còn sử dụng các luật để tạo khoảng cách trắng giữa các từ trong câu như sau:

```
:{p,pp,^blk:+blk:null}"[],blk:null:null";
:{v,vt,^blk:+blk:null}"[],blk:null:null";
:{n,na,^blk:+blk:null}"[],blk:null:null";
:{e,^blk:+blk:null}"[],blk:null:null";
```

1. → *Câu tiếng Việt đầu ra tương đương biểu thức UNL "tôi gửi thư ở bưu điện".*

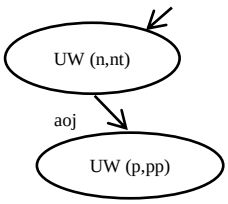
IV. THỬ NGHIỆM VÀ ĐÁNH GIÁ

A. Thử nghiệm

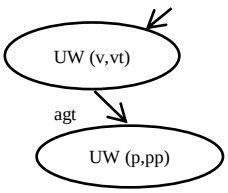
2. Dữ liệu thử nghiệm

Dữ liệu thử nghiệm là các biểu thức chứa 4 quan hệ "*aoj*", *"agt", "obj" và "plc"* với quan hệ nhị phân:

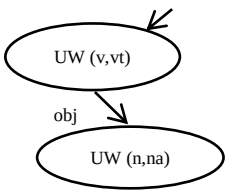
- aoj(n, nt; p, pp)



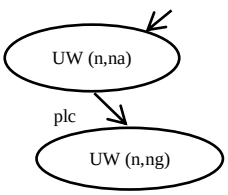
- agt(v, vt; p, pp)



- obj(v, vt; n, na)



- plc(n, na; n, ng)



3. Công cụ DeCovie

Dựa trên giải pháp đề xuất, chúng tôi xây dựng công cụ DeCovie gồm có bốn cửa sổ: cửa sổ thứ nhất chứa các tập luật giải hóa, cửa sổ thứ hai chứa bộ từ điển UNL - tiếng Việt, cửa sổ thứ ba là vùng nhập một biểu thức UNL, cửa sổ thứ tư là trình bày quá trình tìm và thực hiện lệnh.

Hình 5. Cửa sổ phân tích luật của công cụ DeCovie

4. Kết quả thử nghiệm

- Chúng tôi thử nghiệm 200 biểu thức UNL để dịch từ biểu thức UNL sang câu tiếng Việt bởi công cụ DeCovie. Sau đó chúng tôi so sánh kết quả của hai công cụ EUGENE và DeCo [9], câu tiếng Việt được tạo ra ở ba công cụ hoàn toàn giống nhau về cấu trúc và nghĩa của câu.

- Chúng tôi tiến hành một thử nghiệm thứ hai, lần thử nghiệm này sử dụng 200 câu tiếng Anh để dịch sang UNL bởi công cụ UNL.ru. Kết quả thử nghiệm được so sánh một công cụ dịch trực tiếp từ tiếng Anh sang tiếng Việt được thống kê trong bảng 1 như sau:

Bảng 1. Tỷ lệ thay đổi giữa dịch qua UNL và trực tiếp

Cách thức dịch	Số câu thay đổi	
	Không thay đổi	Thay đổi
UNL.ru: Tiếng Anh → UNL DeCovie: UNL → tiếng Việt	123 (61,5 %)	77(38,5 %)
Google: tiếng Anh → tiếng Việt	118 (59 %)	82 (41%)

B. Đánh giá

Công cụ DeCovie cho kết quả dịch giống như các công cụ DeConverter khác. Tuy nhiên, so với EUGENE và DeCo thì công cụ DeCovie có ưu điểm:

+ Quá trình phân tích cú pháp không phụ thuộc vào từ điển.
+ Không sử dụng nhiều luật thay thế khi phân phân tích cú pháp.
+ Định nghĩa ít loại luật.

Nhưng DeCovie vẫn có một số hạn chế:

- Chưa định nghĩa đầy đủ các mối quan hệ nhị phân có thể có của UNL.
- Chưa định nghĩa hết những trường hợp có khả năng giữa hai UW.

Vì DeCovie xây dựng giới hạn trên một số quan hệ nên khi kết hợp với công cụ UNL. Ru để dịch từ tiếng Anh sang tiếng Việt có tỷ lệ câu bị thay đổi khá cao (38,5 %) so thử nghiệm riêng công cụ. Nguyên nhân là do:

- Trong tiếng Anh, động từ "*to be*" khẳng định về một vấn đề nào đó (*ví dụ nghề nghiệp*) thì tương đương tiếng Việt với hệ từ "*là*". Nhưng khi chuyển câu "*I am teacher*" khi từ công cụ UNL.ru sang UNL thì "*am*" là động từ "*to be ở thì hiện tại*" nhưng công cụ DeCovie thì bỏ qua động từ "*tobe*" vì chưa định nghĩa trong hệ thống nên câu tiếng Việt tạo ra vẫn mang hàm ý nghĩa câu gốc nhưng sai về ngữ pháp vì thiếu hệ từ "*là*" thành câu "*tôi giáo viên*".

- Hoặc định nghĩa về địa điểm nơi chốn thì cả hai công cụ đều sử dụng quan hệ ngữ nghĩa "*plc*", tuy nhiên công cụ UNL.ru sử dụng cấu trúc `plc(do,thing)` trong khi công cụ DeCovie sử dụng cấu trúc `plc(thing,thing)` để dịch nên kết quả đầu ra chưa chính xác về giới từ chỉ nơi chốn.

V. KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày phương pháp chuyển đổi một biểu thức UNL có một hay nhiều nút con sang câu tiếng Việt tương ứng. Chúng tôi xây dựng một công cụ DeCovie để chuyển đổi từ biểu thức UNL sang câu tiếng Việt gồm các nội dung: phân tích cú pháp biểu thức UNL và câu đơn tiếng Việt để xây dựng luật giải mã, giải pháp liên kết mục từ của từ điển UNL - tiếng Việt.

Công cụ DeCovie được xây dựng cho tiếng Việt nên có nhiều ưu điểm hơn so với các công cụ khác, tuy nhiên chúng tôi mới xây dựng cho hai trường hợp biểu thức UNL cho một số quan hệ nhị phân khi chuyển đổi sang tiếng Việt. Kết quả của bài báo sẽ là cơ sở để chúng tôi nghiên cứu phát triển công cụ DeCovie hoàn chỉnh cho tiếng Việt. Trong tương lai, chúng tôi sẽ tiếp tục phát triển cho các trường hợp khác của biểu thức UNL chứa một hay nhiều biểu thức cho tất cả các quan hệ được định nghĩa trong UNL.

VI. 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, 2015.

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

[3] UNL centre, "DeCoverter Specifications", Version 2.7, http://www.undl.org, 2002.

[4] P. KUMAR and R. K. SHARMA, "Punjabi DeConverter for generating Punjabi from Universal Networking Language", Journal of Zhejiang University-SCIENCE C (Computers & Electronics) ISSN 1869-1951 (Print); ISSN 1869-196X (Online), pages 179 - 196, 2013.

[5] A. K. Saha, M. F. Mridha and J. K. Das, "Semantic Analysis of Bangla Language for Developing A UNL Deconverter", International Journal of Advanced Research in Computer Science and Software Engineering, pages 273-278, 2012.

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

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

[8] P. T. L. Thuyen and V. T. Hung, "Expand data on UNL - Vietnamese dictionnary of UNL Explorer", Journal of Science and Technology, University of Danang, No.56, 2014.

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

[10] P. T. L. Thuyen, V. T. Hung, "Automatic translation of Vietnamese simple sentences based on UNL", 3rd National Foundation for Science and Technology Development Conference on Infornation and Computer Science, pages 218 - 222, 2016.

[11] P. T. L. Thuyền, V. T. Hùng, "Phân tích động từ và câu ghép tiếng Việt trong hệ thống dịch máy dựa trên UNL", Kỷ yếu Hội thảo quốc gia lần thứ XIX: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông, trang 287 - 292, 2016.

[12] Dự án VLSP, "Xử lý văn bản tiếng Việt", http://vlsp.vietlp.org

[13] Đ. B. Lâm, L. T. Hương, "Cải tiến giải thuật Earley trong phân tích cú pháp tiếng Việt", Kỷ yếu hội thảo ICTrda08- VLSP, 2008.

[14] Đ. Điền, "Giáo trình xử lý ngôn ngữ tự nhiên", Nhà xuất bản Đại học Quốc gia Tp Hồ Chí Minh, 2006.

[15] P. T. L. Thuyền, V. T. Hùng, "Rút trích biểu thức UNL để tạo từ điển tiếng Việt - UNL", Tạp chí Khoa học và Công nghệ các trường kỹ thuật, số.110, trang 86 - 90, 2016.

[16] http://www.unl.ru/deco.html.

[17] http://www.unlweb.net.

[18] https://vlsp.hpda.vn/demo/?page=seg_pos_chunk.

SYNTAX ANALYSIS OF VIETNAMESE SINGLE SENTENCES TO BUILDING DECONVERTER TOOL

Phan Thi Le Thuyen, Vo Trung Hung

***ABSTRACT**: The Universal Networking Language (UNL) has been used by many researchers as an intermediate language in automatic translation. Automatic translation system based on UNL consists of two components: EnConverter and DeConverter. DeConverter is used to create a natural language from an UNL expression based on the UNL - Natural language dictionary and decode rules. In this paper, we present a method for converting a UNL expression with one or more child nodes into a corresponding Vietnamese sentence is called DeCovie. We carry out a grammatical analysis between the UNL expressions with one or more child nodes and single Vietnamese sentences and propose two rules type to convert from UNL expressions to Vietnamese single sentences. We use 200 UNL expressions to convert to Vietnamese, which translates well to other Deconverter tools.*