International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 03, March-2015

Application UNL Tools for Vietnamese

Phan Thi Le Thuyen
University of Science and Technology
The University of Danang
Danang, VietNam

Vo Trung Hung
University of Science and Technology
The University of Danang
Danang, VietNam

***Abstract* — In the recent years, Universal Networking Language (UNL) has been interested by researchers in the field of natural language processing. The UNL is an artificial language used to represent information that is independent with natural languages. The main purpose of UNL is to allow people in the world can access information on the Internet in their own language. Currently, many projects are researching to apply UNL in different languages except Vietnamese. In this paper, we introduce the tools that are UNL applications and how to reuse its for the process of encoding a Vietnamese sentence into UNL expression and decoding an UNL expression into Vietnamese. The testing tools are implemented by a number of different examples.**

Keywords* — *UNL system; Universal Networking Language (UNL); IAN; EnCo; EUGENE; Deco

I. INTRODUCTION

Internet is the largest warehouse of documents and information, but how can we exploit it without language barriers is controversial. According to the statistics on http://www.internetworldstats.com, English is the language that is used most on the site. However, not all Internet users know English, so the language barrier is still a major obstacle to exploit the content. Moreover, more and more content on the Internet written in other languages such as Chinese, Japanese, Spanish, etc. The proportion of those who know these languages are far less than that of the English. To avoid the language barrier, one of the solutions is often used is to provide multilingual software or website written by multi-language. However, the cost for this increases much.

An effective solution is automatically translated into the language that users can read and understand. However, there are more than 5,000 languages, so the development of automated translation software for each language pair is extremely difficult, especially for the languages that have less number of users [2]. So, one of the solutions to translate a text from the source language into the target language is to use an intermediate language and to develop an automated translation software from any language into the intermediate language. Thus, with n language instead of having n*(n-1) translation pairs, we just have 2*n translation pairs [6].

Vietnam has 54 ethnic groups and peoples are speaking in their native language. The development of automatic translation for all language pairs is extremely difficult (2862 language pairs). Therefore, the need to study another solution to develop the translation services for all languages.

UNL (Universal Networking Language) is one of the intermediate languages and presently translation through UNL has been developed for about 50 different languages [3]. For example, research results for Russian and English at http://www.unl.ru; UNL Platform tools for 48 languages at the address in the address www.undl.org/unlpf or http://www.unl.ru, www.undl.org/unlexp, http://www.eolss.net/,... In addition, the tool for supporting the independent encoding and decoding process of language provided by UNDL at http://www.undl.org/ and http://www.unlweb.net/wiki /tools.

Although the UNL has potential and great application, in Vietnam but UNL researches are limited. Our team has been carrying the research on the UNL since 2005 and has achieved some results such as: studying possibilities to apply UNL for Vietnamese [6], building UNL-Vietnamese dictionary [3] [4], testing tools available for UNL to apply for Vietnamese language. In this paper, first we briefly introduce UNL language and system. Second, we introduce some tools that we have developed in the research project related to UNL. The final section presents the experimental tools for Vietnamese language. We especially would like to mention translating a sentence from Vietnamese language into UNL and vice versa. From the experimental results, we show the orientation for further research that can be applied UNL in developing automatic translation software for Vietnamese and ethnic minority languages in Vietnam as Cham, Khmer, Co-tu, …

II. UNL LANGUAGE AND SYSTEM

The idea of developing multilingual translation system is using a language as a pivot language and the language can cover the whole content of any natural language. In 1996, UNL was proposed by Dr. Hiroshi Uchida in Senior Research Institute, United Nations University, Tokyo, Japan [1]. UNL has all the corresponding components of a natural language. It includes word concept expression called Universal Words (UWs). The UWs linked together to generate a UNL expression of a sentence. These links are called Relations, which define the role of each word in a sentence. To show the speaker's point of view is expressed through the Attributes.

- *Universal Words*: Although UWs mainly used English words, UWs has some words of other languages and the semantic information to define the concept of natural language. In this way, it allows to limit the inherent ambiguity of words in natural language. For example, the English word "state" will have two different words: state (icl> country) to represent the country; state (icl> region) to denote a region of the country.

- *Relations*: There are 56 relations in UNL, which are used to connect two UWs to build semantic network of a UNL expression. These relations are the edges of the UNL graph or binary relations which directly generate UNL expressions.

- *Attributes*: attributes used for the purpose of describing the subjective information of sentence. They show speaker's point. There are 87 attributes to clarify the semantics of the sentence.

In addition, the UNL system also uses the knowledge base so that it will provide semantic definition of the concept and confirm the extent of relations to avoid definite ambiguity.

For example, the sentence "I can hear a dog barking outside" represented by the UNL language as follows:

```
{unl}
agt(hear(icl>perceive(agt>person,obj>thing))
.@entry, I)
obj(hear(icl>perceive(agt>person,obj>thing))
.@entry, :01)
agt:01(bark(agt>dog).@entry,
dog(icl>canine))
plc:01(bark(agt>dog).@entry,outside(icl>plac
e)
{/unl}
```

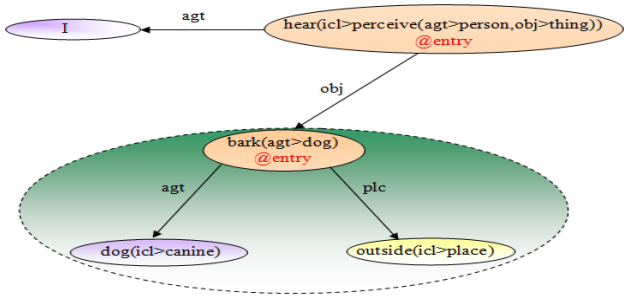

Figure 1. A sentence are expressed in terms of a graph

To integrate it into the general UNL system, each natural language just builds a language server on the Internet with Enconverter and Deconverter functions. Enconverter function is responsible for converting documents from natural language into the UNL language and Deconverter function translates the text from UNL language into natural language (NL). For the encoding and decoding process, we both have to build a set of encoding and decoding rules and need a bilingual UNL – NL dictionary.

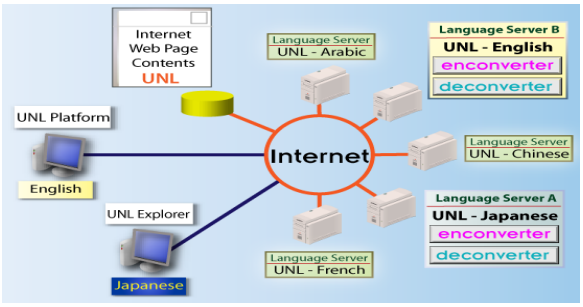

Figure 2. UNL multilingual translation system

III. SOME SUPPORTING TOOLS FOR UNL

A. *Tools for Enconverter*

To perform the encoding process, we can use a number of available tools such as:

1) IAN tool

IAN (Interactive Analyzer) is software developed on the Web environment to perform the encoding [7]. Each natural language can be integrated and stored on the server. Each natural language can be integrated and stored on the server. Therefore, data sources can be exploited anytime, anywhere without being dependent on geographical distance. IAN analyzes the input with the help of T-Rules and dictionaries. IAN has 8 tags: Welcome, NL Input, Dictionaries, N-Rules, T-Rules, D-Rules, IAN and Compare.

The tag "NL Input" allows users to provide documentation of natural language. The tag "Dictionaries" allows users to provide dictionaries NL - UNL according to the specifications of the UNL [9]. The tag "T-Rules" allows users to provide rules of grammar conversion from natural language into UNL. The tag "D-Rules" provides grammar orientation from natural language into UNL. This tag is used to control token and improve the result of grammar transformation. The tag "IAN" shows the encoding result prescribed by UNL. The tag "Compare" is the tag to compare the results with other results.

2) *EnCo tool*

The EnCo is a tool used on a single machine [7]. Each of the input string is scanned from left to right and combined taking in the word dictionary become the candidate. The candidates are sorted in order of priority. Selecting candidates is done by applying the rules of grammar. The parsing and semantics is done by applying the rules to choose words to build a syntax tree and a semantic network for the input. This process continues until all of the words in a sentence were put into and a complete semantic network of the input. The output of the entire process is a semantic network presented in UNL format.

B. *Tools for Deconverter*

To perform the decoding process, we can use a number of available tools follows as:

1) EUGENE tool

EUGENE (dEp-to-sUrface GENErator) is an application developed on the web environment to create

sentences of natural language from UNL expression based on UNL - NL dictionary and a set of transformation rules [8]. EUGENE has 7 tags: Welcome, NL Input, Dictionaries, T-Rules, D-Rules, EUGENE and Compare.

The tag "NL Input" allows users to provide document action in the form of UNL expressions. The tag "Dictionaries" allows users to supply UNL-NL dictionary according to the specifications of the UNL [10]. The tag "T-Rules" provides grammar conversion from UNL into NL. The tag provides grammar orientation to convert from UNL into natural language. This tag is used to control "*token*" and improve grammar transformation result. The tag "EUGENE" shows that the result of encoding is complete sentences in natural language. The tag "Compare" is the tag to compare the results with others.

2) *DeCo tool*

Deco tool [10] works as follows:

First, transforming a set of binary relation of UNL expressions input into a directed graph with the hyper-nodes called node-net. The root node of node-net called entry node and starts at the beginning of each sentence. Then applying the decoding rules to node of the node-net. It begins with "entry node" to find an appropriate word for each node and create a word string in grammatical order of natural language. The decoding process ends when all the words to all the nodes are found and a word string of target language is completed.

IV. TESTING FOR VIETNAMESE

A. Enconverter

1) IAN tool (Interactive Analyzer)

- Input: "Quyển sách viết về Sài gòn"

- After analyzing the above sentence, we have the following result:

[Quyển sách] [ ] [viết] [ ] [về] [ ] [Sài gòn]

- The Vietnamese - UNL dictionary for above sentence will be:

```
[quyển sách]{}"book(icl>publication>thing)"
(LEX=N,POS=NOU,NUM=SNG)<vie,0,0>;

[viết]{}"write(icl>communicate>do,agt>person
,obj>information,cao>thing,ins>thing,ec>pers
on)"(LEX=V,POS=VER,PER=3PS)<vie,0,0>;

[về]{}"about"(LEX=P,POS=PRE,rel=PLC)<vie,0,0
>;

[sài gòn]{}"saigon(iof>city>thing,equ>
ho_chi_minh_city)"(LEX=N,POS=PPN,NUM=SNGT)<v
ie,0,0>;
```

- Use rules to encode for the above sentence:

Rule 1: Remove the blank "[ ]"

(%a,BLK):=;

Rule 2: Delete %b and put attribute into %a

(P,PRE,att=@about,%b)(N,PPN,%a):=(%a,+att=%b);

Rule 3: Assign attributes @and @present into verb node

(V,VER,3PS,^@entry,%a):=(%a,+@entry,+@present);

Rule 4: Assign attribute @def into noun node

(N,NOU,^@def,%a):=(%a,+@def);

Rule 5: Create a new button "NA" that has attribute (N, AGT) and NB has attribute (N,CAO)

(N,NOU,%a)(V,VER,%b)(N,PPN,%c):=(NA(%b;%a),+ N,+AGT) (NB(%b;%c), +N, + CAO);

Rule 6: change relation "NA" to "agt" keep content of root node.

(NA(%a;%b),AGT):=agt(%a;%b);

- Output

```
### IAN Console ###

2. Pattern: [quyển sách] [ ] [viết] [ ] [về] [ ] [Sài gòn]

 [S:2]
 {org}
 quyển sách viết về Sài gòn
 {/org}
 {unl}
 agt(write(icl>communicate>do,agt>person,obj>information,cao>thing,ins>thing,rec>person):03.@entry.@present,
book(icl>publication>thing):01.@def)
 cao(write(icl>communicate>do,agt>person,obj>information,cao>thing,ins>thing,rec>person):03.@entry.@present,
saigon(iof>city>thing,equ>ho_chi_minh_city):07.@about)
 {/unl}
 [/S]
Dictionary Lookup Time 0 seconds, 13 milliseconds.
Tokenization Time 0 seconds, 0 milliseconds.
Transformation Time 0 seconds, 10 milliseconds.
Total Time 0 seconds, 23 milliseconds.
```

Figure 3. UNL expression are created correspondence

2) *EnCo tool*

- Input: "Tôi ăn cơm"

- The Vietnamese - UNL dictionary for above sentence will be:

```
[tôi]{}"I(icl>person)"(PRON,HPRON,1SG,SUBJ)<
v,0,0>;

[cơm]{} "rice(icl>food)"(N,SUBJ)<v,0, 0>;

[ăn]{}"eat(icl>consume>do,agt>living_thing,o
bj>concrete_thing)" (VT, S, 3SG.S,AGT.S,BA,
BAE, OBJ.S,V,VDO,OBJ.S)<v,0,0>;
```

- Use rules to encode for the above sentence:

Rule 1: Move the window to the right

R {}{^mor,^STAIL:mor}P200;

Rule 2: Change the attribute "mor" for STAIL (<<)

: {}{^mor,STAIL:mor}P200;

Rule 3: Move the window to the right

R {}{mor}P1;

Rule 4: Combine with a blank on left

+ {^blk:blk}{BLK}P210;

Rule 5: Add EONP and HNP at the end of the noun

: {}{mor,N,SUBJ,^EONP:EONP,hnp}P150;

Rule 6: Remove a blank on the right of the noun

DR {N,PN,PRON}{BLK}P200;

Rule 7: Create relation between verbs and nouns

< {V,VDO,OBJ.S,^psv,^>obj}{N,EOR::obj}P100;

Rule 8: Add tense and form into the verb

:{}{mor,V,S,^EN,^&@present.@entry:-s, BA , & @present .@entry,s}P90;

Rule 9: There is a semantic relation "AGT" between "tôi" and "ăn". Apply the rule

>{PRON,SUBJ,^prep:&@topic:agt}{V,VT,AGT.S,^>agt, ^psv,^subj:subj} P100;

Rule 10: There is a semantic relation "obj" between "cơm" and "ăn". Apply the rule

>{N,SUBJ,^prep:&@topic:obj}{V,VT,OBJ.S,^>obj,^psv , ^subj:subj}P100;

Rule 11: Add EOR at the right end

: {^EOR:EOR}{STAIL}P200;

- Output:

```
[S:1]
{org}
tôi ăn cơm
{/org}
{unl}
[W]
:00
[/W]
{/unl}
{unl}
obj(eat(icl>consume>do,agt>living_thing,obj>concrete_thing):08.@present.@entry,        rice(icl>food):0C)
agt(eat(icl>consume>do,agt>living_thing,obj>concrete_thing):08.@present.@entry,        I(icl>person):03.@topic)
{/unl}
[/S]
```

Figure 4. UNL expression are created correspondence

B. Deconverter

1) EUGENE tool

- Input: the beautiful car

- Encoding the English sentence into UNL expression as follows:

```
{unl}
mod(car(icl>motor_vehicle>thing).@entry.@def
,
beautiful(icl>adj,ant>ugly))
{/unl}
```

- The Vietnamese - UNL dictionary for above sentence will be:

```
[xe ô tô] {} "car(icl>motor-vehicle>thing)"
(LEX=N, POS=NOU, NUM=SNGT)<vie,0,0>;
[đẹp]{}"beautiful(icl>adj,ant>ugly)"(LEX=J,P
OS=ADJ)<vie,0,0>;
```

- Use rules to encode for the above sentence:

Rule 1: Remove the relation "mod", create new relation "NA"

mod(%x,N;%y,J):=NA(%x;%y,DIS=BEF);

Rule 2: Remove the attribute @def, create new relations "NS" with two UWs "%x" và "chiếc" with attribute is (LEX = D, + POS = ART)

(%x,N,@def):=(NS(%x,@def;%y,[chiếc],+LEX=D,+POS =ART));

Rule 3: Create blank between the UWS in the relation "NA"

NA(%x;%y):=(%x,+>BLK)(%y,+>BLK);

Rule 4: Create blank between the UWS in the relation "NS"

NS(%x;%y):=(%y,+>BLK)(%x,+>BLK);

Rule 5: Create blank between words in a sentence

(%x,>BLK)(%y,^BLK,^PUT,^STAIL):=(%x,->BLK)("",+BLK)(%y);

- Output :

```
### Please select the sentence to run from the list above ###

 [S:1]
 {org}
 the beautiful car
 {/org}
 {vie}
 chiếc xe ô tô đẹp
 {/vie}
 {unl}
 mod(car(icl>motor-vehicle>thing).@entry.@def,beautiful(icl>adj,ant>ugly))
 {/unl}
 [/S]
Dictionary Lookup Time 0 seconds, 4 milliseconds.
Disambiguation Time 0 seconds, 0 milliseconds.
Transformation Time 0 seconds, 12 milliseconds.
Total Time 0 seconds, 16 milliseconds.
```

Figure 5. UNL expression are created correspondence

2) *DeCo tool*

- UNL Expression input:

```
[S:1]
{org}
I ate rice
{/org}
{unl}
agt(eat(icl>consume>do,agt>living_thing,obj>
concrete_thing).@entry.@past,I(icl>
peson))
obj(eat(icl>consume>do,agt>living_thing,
obj>concrete_thing).@entry.@past,rice(icl>fo
od))
{/unl}
[/S]
```

- The Vietnamese - UNL dictionary for above sentence will be:

```
[tôi]{}
"I(icl>person)"(PRON,HPRON,1SG,SUBJ)<v,0,0>;
[cơm]{} "rice(icl>food)"(N,SUBJ)<v,0, 0>;
[đã
ăn]{}"eat(icl>consume>do,agt>living_thing,
obj>concrete_thing)" (VT, S, 3SG.S,AGT.S,BA,
BAE, OBJ.S,V,VDO,OBJ.S) <v,0,0>;
```

- Use rules to encode for the above sentence:

Rule 1: Move the window to the right

R {}{} P1;

Rule 2: Insert object of relations "AGT" into Node List

:"HPRON,SUBJ,1SG:subj:agt"
{V,^IRG,^pred:pred, 1sg} P120;

Rule 3: Insert object of relations "obj" into Node List

: {V,VDO,pred,^OBJ_inserted:OBJ_inserted}
":obje: obj" P100;

Rule 4: Insert a blank button for pronouns

: {PRON,^blk:blk} "[ ]:blk" P80;

Rule 5: Insert a blank button for verb

: {V,^blk:blk} "[ ]:blk" P80;

```
[S:1]
================== WORD LISTS : 03 ==================
01  00  00:<agt
[tôi]{}  "I(icl>person)" (PRON,HPRON,1SG,SUBJ) <v,0,0>;
02  00  00:@entry,@past,>agt,>obj
[đã ăn]{}  "eat(icl>consume>do,agt>living_thing,obj>concrete_thing)" (VT,S,3SG.S,AGT.S,BA,BAE,OBJ.S,V,VDO,OBJ.S) <v,0,0>;
03  00  00:<obj
[cơm]{}  "rice(icl>food)" (N,SUBJ) <v,0,0>;
======================== UNL ========================
eat(icl>consume>do,agt>living_thing,obj>concrete_thing)(@entry,@past,>agt,>obj)
  -agt>I(icl>person)(<agt)
  -obj>rice(icl>food)(<obj)
=============================================================
tôi đã ăn cơm
;;Time  0.0          Sec
;;Done!
```

Figure 6. UNL expression are created correspondence

C. Evaluate

EnCo and Deco have interface as well as the encoding results that are unfriendly with users and software used on a single computer, it is difficult to share data for other users. IAN and EUGENE are built on platforms web and friendly interface. Special resources like: dictionaries, rules can be shared to other users in the online community.

We have tested the tool with some Vietnamese sentences such as: encoding from a Vietnamese sentence into UNL and from UNL into Vietnamese with very satisfactory outputs about the translation quality. We will continue to expand the UNL - Vietnamese dictionary, and share in the online community.

V. CONCLUSION

Through experimental results, we can find that converter tools from natural language to UNL and vice versa are very effective. The quality is pretty good and acceptable. However, to use these tools, users must comply with the specified data format. But however, there still have some languages not follow the rules, so these tools need adjusting.

In the coming time, we will install a Vietnamese automatically translated server with two functions: Enconverter and Deconverter. In addition, in the next stage, we will continue to install the new server for the languages of the ethnic minorities in Vietnam such as Chăm, Khmer, Cơ-tu, etc In order to have a complete multilingual automatic translation system for the languages being used in Vietnam.

REFERENCES

[1] H. Uchida (1996), "UNL: Universal Networking Language – An Electronic Language for Communication, Understanding, and Collaboration", UNU/IAS/UNL Center, Tokyo, Japan.

[2] S. Tripathi and J. K. Sarkhel (2010), "Approaches to machine translation", Annals of Library and Information Studies, vol. 57, pages 388-393.

[3] Phan Thị Lệ Thuyền, Võ Trung Hùng (2014), "Bổ sung dữ liệu vào từ điển UNL – tiếng Việt trong bộ công cụ UNL Explorer", Tạp chí Khoa học Công nghệ ĐHĐN, Số 11(84).2014, Quyển 1.

[4] V. T. Hung, G. Fafiotte (2011), "UVDict – a machine translation dictionary for Vietnamese language in UNL system", Proceeding CISIS 2011, Korean Bible University (KBU), Seoul, Korea, Pages: 1020-1028.

[5] V. T. Hung (2004), "Reuse of Free Online MT Engines to Develop a Meta-system of Multilingual Machine Translation", EsTAL 2004, 2004.

[6] V. T. Hùng (2007), "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.

[7] P. Kumar, R.K. Sharma (2012),"Punjabi to UNL EnConversion System", Springer: Sadhna, Academy Proceedings in Engineering Sciences, volume 37:(2), pp 299–318.

[8] Shi X, Chen Y, (2005), "A UNL DeConverter for Chinese Universal Network Language", Universal Network Language: Advances in Theory and Applications, Ed(s) Cardenas J, Gelbukh A, Tovar E, México, Research on Computing Science: 167-174.

[9] UNL centre (2002), Enconverter Specifications, Version 3.3, http://www.undl.org/.

[10] UNL centre (2002), Deconverter Specifications, Version 2.7, http://www.undl.org/.