BỔ SUNG DỮ LIỆU VÀO TỪ ĐIỂN UNL – TIẾNG VIỆT TRONG BỘ CÔNG CỤ UNL EXPLORER

EXPANSION OF UNL – VIETNAMESE DICTIONARY ON UNL EXPLORER

Phan Thị Lệ Thuyền, Võ Trung Hùng

Trường Đại học Bách khoa, Đại học Đà Nẵng; Email: thuyenptl@gmail.com, vthung@dut.udn.vn

Tóm tắt - Một dự án nghiên cứu thu hút nhiều nhà khoa học, tổ chức và cá nhân là phát triển hệ thống UNL. Một trong những khâu quan trọng khi phát triển hệ thống UNL là xây dựng bộ từ điển của mỗi ngôn ngữ và tích hợp vào hệ thống. Trong bài báo này, chúng tôi đề xuất giải pháp mở rộng một từ điển UNL – Tiếng Việt thông qua việc sử dụng công cụ UNL Explorer và các công cụ tự phát triển. Phương pháp chúng tôi sử dụng là trích tự động các mục từ trong các từ điển Anh - Việt để đối chiếu với các mục từ có sẵn của UNL – Tiếng Anh, nếu mục từ nào chưa tồn tại thì chúng tôi bổ sung mục từ tiếng Việt tương ứng vào từ điển UNL - Tiếng Việt. Đối với những mục từ còn thiếu thì chúng tôi nhập thủ công bằng công cụ UNL Explorer. Kết quả đạt được là chúng tôi bổ sung thêm được 30.000 từ mới vào từ điển UNL – Tiếng Việt và nhập mới thêm 550 từ bằng thủ công.

Từ khóa - dịch máy, hệ thống UNL, ngôn ngữ UNL, từ điển, từ điển UNL – Tiếng Việt.

Abstract - A research project that has attracted scientists, organizations and individuals around the world is the development of UNL system. One important step in the system development is building dictionaries for all languages and integrating them into the UNL system. In this paper, we propose a solution to expand the UNL - Vietnamese dictionary by using UNL Explorer and other builder tools. Our method is to extract automatically entries from English - Vietnamese dictionary and compare them with the available items in UNL- English. If the item does not exist, we will add it into UNL – Vietnamese dictionary. For those missing entries, we entered manually by UNL Explorer tool. As the result, we added 30,000 new words into the dictionary UNL - Vietnamese and 550 new words were entered manually.

Key words - machine translation, UNL system, Universal Networking language (UNL), dictionary, UNL – Vietnamese dictionary.

1. Giới thiệu

Internet đã trở nên phổ biến và là một trong những kênh cung cấp thông tin lớn nhất hiện nay. Đối tượng người dùng trên Internet rất phong phú và sử dụng nhiều ngôn ngữ khác nhau. Theo thống kê của **W3Techs** vào tháng 12/2013, về nội dung trên web theo ngôn ngữ của 10 ngôn ngữ phổ biến nhất là:

Bảng 1. *Thống kê nội dung web dựa trên ngôn ngữ*

STT	Ngôn ngữ	Tỉ lệ
1	Tiếng Anh	55.7%
2	Tiếng Nga	6.0%
3	Tiếng Đức	6.0%
4	Tiếng Nhật	5.0%
5	Tiếng Tây Ban Nha	4.6%
6	Tiếng Pháp	4.0%
7	Tiếng Trung	3.3%
8	Tiếng Bồ Đào Nha	2.3%
9	Tiếng Ý	1.8%
10	Tiếng Ba Lan	1.7%

Trong khi đó, thống kê về số lượng người dùng trên Internet như sau:

Bảng 2. *Số lượng người dùng theo ngôn ngữ*

STT	Ngôn ngữ	Tỉ lệ
1	Tiếng Anh	27%
2	Tiếng Trung	25%
3	Tiếng Tây Ban Nha	8%
4	Tiếng Nhật	5%
5	Tiếng Bồ Đào Nha	4%
6	Tiếng Đức	4%
7	Tiếng A-rập	3%
8	Tiếng Pháp	3%
9	Tiếng Nga	3%
10	Tiếng Hàn	2%

Ngoài ra, người dùng và nội dung có trên Internet đang sử dụng hàng trăm ngôn ngữ khác nhau. Vậy vấn đề đặt ra là làm thế nào để người sử dụng có thể trao đổi với nhau hoặc khai thác được những nội dung viết trong những ngôn ngữ mà họ không biết?

Để phá vỡ rào cản về ngôn ngữ, những giải pháp thường được sử dụng hiện nay là đa ngữ hóa các hệ thống (nhằm cho phép người dùng lựa chọn ngôn ngữ sử dụng đối với phần mềm/website) hoặc hỗ trợ người dùng thông qua các phần mềm dịch tự động.

Một trong những hệ thống hỗ trợ đa ngữ hóa và dịch tự động được quan tâm nghiên cứu hiện nay là UNL (Universal Networking Language). Mục đích chính của hệ thống UNL là cung cấp cho người sử dụng Internet truy cập vào các trang web trong ngôn ngữ mà họ lựa chọn. Hiện nay, nhiều ngôn ngữ (45 ngôn ngữ vào cuối năm 2013) đã được tích hợp vào nền tảng của UNL như: Tiếng Anh, tiếng Pháp, tiếng Nga, tiếng Nhật, tiếng Trung, tiếng Tây Ban Nha,…

Nhằm mục đích nghiên cứu về tiếng Việt và tích hợp tiếng Việt vào hệ thống UNL, chúng tôi đã triển khai một số nghiên cứu và đã đạt được một số kết quả ban đầu [1]. Trong bài báo này, chúng tôi tập trung giới thiệu việc mở rộng kho dữ liệu từ điển UNL – Tiếng Việt. Chúng tôi sử dụng các từ điển sẵn có như từ điển UNL – Tiếng Anh (2.080.318 từ), UNL – Tiếng Việt (651.984 từ) của công cụ UNL Explorer và từ điển Anh – Việt của Hồ Ngọc Đức để mở rộng dữ liệu từ điển UNL – Tiếng Việt [4].

Phương pháp chúng tôi sử dụng là trích tự động các mục từ trong từ điển Anh - Việt để đối chiếu với các mục từ có sẵn của UNL – Tiếng Anh, nếu mục từ này chưa tồn tại thì chúng tôi bổ sung mục từ tiếng Việt tương ứng vào từ điển UNL - Tiếng Việt. Đối với những mục từ còn sót

(có trong từ điển UNL - Tiếng Anh mà không có trong UNL – Tiếng Việt) thì chúng tôi nhập thủ công bằng công cụ UNL Explorer.

Kết quả đạt được là chúng tôi bổ sung thêm được 30.000 từ mới vào từ điển UNL – Tiếng Việt dựa trên từ điển Anh – Việt và nhập mới thêm 500 từ bằng thủ công.

Bài báo được tổ chức thành các phần chính như sau: giới thiệu về hệ thống UNL, cấu trúc từ điển UNL, công cụ UNL Explorer, giải pháp và thực hiện việc mở rộng dữ liệu từ điển UNL – Tiếng Việt.

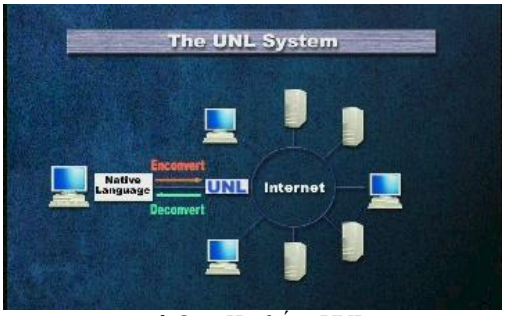
2. UNL và hệ thống UNL

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 tri 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 [6].

UNL bao gồm các thành phần để biểu diễn một ngôn ngữ tự nhiên: UW (Universal Word, kho từ vựng), Relation (Quan hệ), Attributes (Thuộc tính) và UNL Knowledge Base (UNLKB, Cơ sở tri thức).

UNL liên kết các từ vựng dựa trên mô tả quan hệ và thuộc tính để tạo thành câu. Những liên kết này gọi là "relation", nó chỉ định vai trò của mỗi từ trong câu và ngụ ý của người nói được diễn tả thông qua "attribute". UNLKB định nghĩa quan hệ có thể có giữa các khái niệm, bao gồm các quan hệ phân cấp và kỹ thuật tham chiếu giữa các khái niệm. Vì thế, UNLKB cung cấp nền tảng ngữ nghĩa của UNL để chắc chắn nghĩa của biểu thức UNL là không nhập nhằng.

Để phát triển một hệ thống chuyển đổi từ tiếng Việt → UNL và ngược lại cần hai công cụ chính là mã hóa (EnConverter) [2] và giải mã (DeConverter) [3]. Công cụ mã hóa thực hiện một phân tích cú pháp ngôn ngữ độc lập về hình thái, cú pháp và ngữ nghĩa. Công cụ giải mã thực hiện độc lập để chuyển đổi biểu thức UNL sang câu của một ngôn ngữ tự nhiên, nó bao gồm cấu trúc hình thái, cú pháp và lựa chọn từ.



***Hình 1**. Hệ thống UNL*

3. Cấu trúc từ điển unl

Từ điển UNL hay cũng được gọi là từ điển tổng hợp các khái niệm của UNL (Dictionary of UNL Concepts) là một phần của dự án quốc tế nhằm phát triển hệ thống UNL. Sự phát triển nguồn tài nguyên này được hỗ trợ bởi các thành viên của "U++ Consortium", trong đó bao gồm các nhà nghiên cứu từ Nga, Pháp, Tây Ban Nha, Ấn Độ và một số quốc gia khác [7].

Các đơn vị cơ bản của từ điển được gọi là "khái niệm UNL" (UNL concept), nó tương ứng với nghĩa của từ được mô tả trong các từ điển được xây dựng theo cách truyền thống. Định nghĩa một khái niệm của UNL cũng phù hợp với từ điển truyền thống. Điều này cho phép tái sử dụng nhiều dữ liệu ngôn ngữ tự nhiên đã thu thập được từ các bộ từ điển và các bộ bách khoa toàn thư.

Các mục từ của từ điển UNL được tổ chức theo một cấu trúc nhất định. Cấu trúc chung như sau:

```
[HW]{} "UW" (ATTR1, ATTR2, …) <FLG, FRE, PRI>;
```

Trong đó:

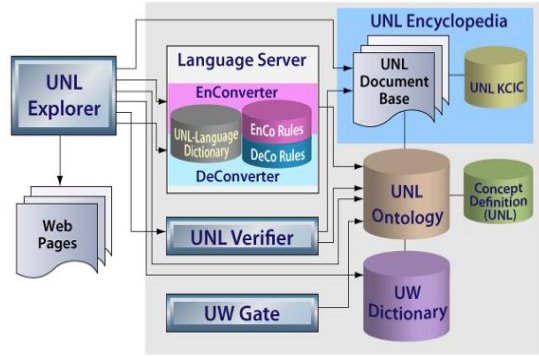
- `HW` (Headword): từ đầu mục từ của ngôn ngữ
- `ID` (Identification): định danh (có thể trống)
- `UW` (Universal Word): từ vựng
- `ATTR` (Attribute): thuộc tính ngữ pháp
- `FLG` (Flag): cờ ngôn ngữ, một ký tự trong bảng mã ASCII
- `FRE` (Frequence): tần số sử dụng trong mã hóa
- `PRI` (Priority): ưu tiên sử dụng trong giải mã

Ví dụ, một mục từ của từ điển UNL – Tiếng Việt được mã hóa như sau:

```
[làm_phát_cáu]{V}
"irritate(agt>human,equ>disturb)"
```

4. Công cụ unl explorer

UNL Explorer là một ứng dụng web dựa trên thông tin đa ngôn ngữ và hệ thống quản lý tri thức. Nó cung cấp cho người dùng một môi trường tích hợp mà ở đó người dùng có thể tìm kiếm và chỉnh sửa các tri thức và thông tin dựa trên UNL mà không bị rào cản bởi ngôn ngữ. Nó quản lý và tìm tri thức và thông tin dựa trên hệ thống từ vựng của UNL Ontology.



***Hình 2.** Cấu trúc của UNL Explorer*

UNL Explorer hỗ trợ các chức năng phục vụ cho người sử dụng thực hiện việc nghiên cứu, phát triển các tài nguyên ngôn ngữ dựa trên UNL như sau: tìm kiếm đa ngữ (Multilingual Context Search); dịch tự động (Translate); từ điển đa ngữ (Multilingual Dictionary); bổ sung bản thể của UNL (UNL Ontology); hỗ trợ giao tiếp (UNL Talk)

Từ điển đa ngữ hiện tại của UNL bao gồm từ điển của 47 ngôn ngữ. Người dùng có thể sử dụng tra cứu cho bất

kỳ cặp ngôn ngữ nào có trong bộ từ điển này (ngôn ngữ tự nhiên – UNL, UNL – ngôn ngữ tự nhiên, ngôn ngữ tự nhiên – ngôn ngữ tự nhiên) [7][8].

Đối với chức năng hỗ trợ từ điển đa ngữ, ta có:

`Word Mean`: Người dùng có thể xem ngữ nghĩa của từ thể hiện qua 47 ngôn ngữ khác nhau.

`Semantic Co-occurrence`: Cho phép xử lý trường hợp có sự xuất hiện đồng thời các từ cùng nghĩa.

`Search`: Cho phép người dùng tìm kiếm các mục từ bằng cách nhập vào mục từ muốn tìm. UNL Explorer sẽ thực hiện tìm kiếm trong từ vựng UNL từ ngôn ngữ được chọn và hiển thị kết quả mục từ được biểu diễn thông qua biểu thức UNL. Người dùng có thể chọn và sửa đổi thông tin cho các mục từ nếu muốn.

`Create New Entry`: Cho phép người sử dụng phát triển thêm các mục từ bổ sung vào cơ sở dữ liệu của UNL Explorer. Đây là một trong những chức năng chính mà UNL Explorer cung cấp để người sử dụng có thể phát triển cơ sở dữ liệu của UNL dành cho bất cứ ngôn ngữ nào được hỗ trợ.

`Delete Entry`: Cho phép xóa mục từ được chọn.

`Show properties`: Cho phép người sử dụng xem định nghĩa liên quan đến mục từ được chọn. Người sử dụng có thể trực tiếp sửa đổi thông tin liên quan đến mục từ ngay trên chính cửa sổ hiển thị để làm tăng độ chính xác và hoàn chỉnh cho cơ sở dữ liệu của UNL Explorer.

`Operation`: cho phép tải về danh sách các UW của từ điển ngôn ngữ đang được chọn hoặc danh sách các UW bao gồm cả ngôn ngữ tự nhiên tương ứng với UW.

Người sử dụng có thể tải các từ điển về để thực hiện nghiên cứu cấu trúc ngữ pháp của ngôn ngữ UNL hoặc phục vụ một số mục đích phát triển khác một cách miễn phí.

5. Xây dựng từ điển unl – tiếng việt

5.1. Từ điển Anh – Việt

Hiện nay, `www.dict.org` đã xây dựng được một định dạng từ điển rất dễ sử dụng, định dạng này đã được một số tổ chức, cá nhân chọn sử dụng để xây dựng những bộ từ điển khá lớn.

Định dạng từ điển `Dict` được mô tả như sau: toàn bộ cơ sở dữ liệu dược chứa trong 2 tập tin dưới dạng văn bản (TXT), một tập tin chứa nghĩa của từ và một tập tin chỉ mục. Tập tin chỉ mục bao gồm tên từ, vị trí nghĩa của từ bắt đầu trong tập tin chứa nghĩa và độ dài của nghĩa. Tập tin chỉ mục được sắp xếp để giảm bớt thời gian tìm kiếm.

Cấu trúc tổng quát của tập tin chứa nghĩa như sau:

```
@ headword
*tu loai (noun, verb...)
-dinhnghia1= cauviduchodinhnghia1+nghiacuacaudo
-dinhnghia2= cauviduchodinhnghia2+nghiacuacaudo
* tu loai
- dinhnghia3
```

Ví dụ: từ "`inside`" trong từ điển Anh – Việt theo chuẩn Dict.

```
@inside /'in'said/
* danh từ
 - mặt trong, phía trong, phần trong, bên trong
 - phần giữa
 + the inside of a week: phần giữa tuần
 - (thông tục) lòng, ruột
 - lộn trong ra ngoài (to turn inside out)
* tính từ và phó từ
 - ở trong, từ trong, nội bộ
 + inside information: tin tức nội bộ
 + an inside job: một công việc làm ở trong
 + inside of a week: trong vòng một tuần
* giới từ
 - ở phía trong; vào trong
 - phần trong, mặt trong, tính chất trong
```

Cấu trúc mục từ của từ điển Anh – Việt của tác giả Hồ Ngọc Đức tuân theo chuẩn `Dict`. Đây là từ điển điện tử được phát hành dưới giấy phép GNU (GPL) và đặt tại http://www.informatik.uni-leipzig.de/~duc/Dict/. Ví dụ cấu trúc mục từ "`abalone`" như sau:

```
@abalone /,æbə'louni/
* danh từ
 - (từ Mỹ,nghĩa Mỹ) bào ngư
@abandon /ə'bændən/
* ngoại động từ
 - bỏm (nhiếp ảnh) (nhiếp ảnh) (từ Mỹ,nghĩa Mỹ) từ bỏ; bỏ rơi, ruồng bỏ
 + to abandon a hope: từ bỏ hy vọng
 + to abandon one's wife and children: ruồng bỏ vợ con
 + to abandon oneself to: đắm đuối, chìm đắm vào (nỗi thất vọng...)
* danh từ
 - sự phóng túng, sự tự do, sự buông thả
 + with abandon: phóng túng
@abandoner /ə'bændənə/
* danh từ
 - (pháp lý) người rút đơn
```

Chúng tôi sử dụng trường `headword` nằm sau ký tự `@` để so sánh với headword của mục từ tiếng Anh trong từ điển UNL – Anh từ UNL Explorer.

5.2. Giải pháp xây dựng từ điển UNL – Tiếng Việt trên UNL Explorer

UNL Explorer đã hỗ trợ sẵn UNL dành cho tiếng Việt với 651.984 từ, UNL dành cho tiếng Anh là 2.080.318 từ. Số lượng từ vựng tiếng Việt hiện có so với số lượng từ vựng tiếng Anh mà UNL Explorer đã xây dựng là quá nhỏ. Chính vì vậy, chúng tôi sử dụng cấu trúc các mục từ của UNL – Tiếng Anh đã được mô tả sang UNL để phát triển và mở rộng thêm các mục từ dành cho UNL – Tiếng Việt.

Qua nghiên cứu cấu trúc từ điển UNL – Tiếng Anh và từ điển Anh - Việt theo chuẩn Dict của tác giả Hồ Ngọc Đức thì chúng tôi nhận thấy rằng, để xây dựng từ điển UNL – Tiếng Việt cần thực hiện các bước sau:

Bước 1: Sử dụng từ điển Anh - Việt của tác giả Hồ Ngọc Đức để tiến hành tổng hợp và chọn lọc các từ vựng.

Bước 2: Tải danh sách các mục từ UNL – Tiếng Việt được phát triển bởi UNL Explorer. Sau đó, chúng tôi tiến

hành so sánh giữa từ vựng trong từ điển Hồ Ngọc Đức và từ vựng đã được xây dựng bởi UNL Explorer, nhằm lựa chọn các mục từ không bị trùng lặp.

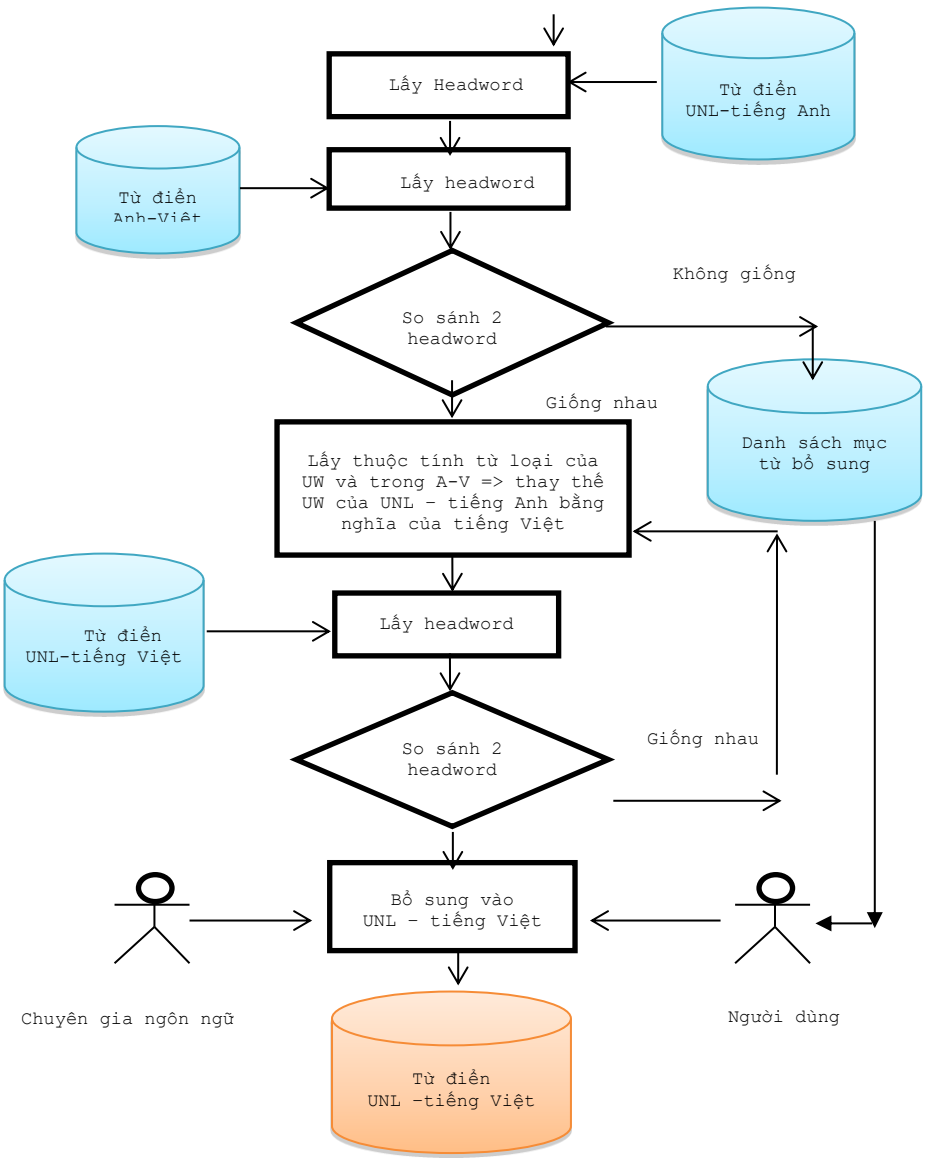
Bước 3: Chúng tôi tiến hành lưu các từ vựng không trùng lặp để bổ sung khoảng 30.000 mục từ.

Bước 4: Sử dụng công cụ UNL Explorer, chúng tôi tiến hành so khớp các Headword trong từ điển của Hồ Ngọc Đức với Headword trong từ điển UNL – Tiếng Anh để trích lọc.

Bước 5: Bổ sung vào từ điển UNL – Tiếng Việt của công cụ UNL Explorer.

5.3. Mô hình hệ thống

Qui trình triển khai như sau:



Hình 3. *Mô hình hệ thống*

Để tạo các mục từ sử dụng cấu trúc của UNL. Chúng tôi thực hiện nghiên cứu dựa trên các thành phần chính của UNL: từ vựng (Universal Words), quan hệ (Relation), thuộc tính (Attribute).

5.4. Thử nghiệm

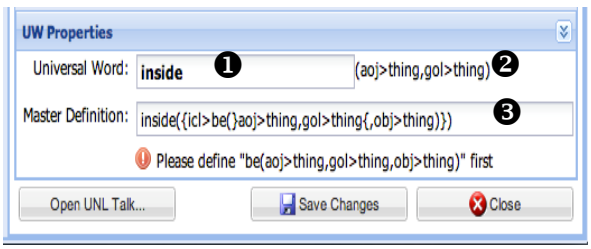
Dữ liệu đầu vào:

- Tập tin chứa danh sách các từ mục được trích từ từ điển của Hồ Ngọc Đức.

- Tập tin chứa các mục từ của từ điển UNL – Tiếng Anh trên UNL Explorer.

Dữ liệu đầu ra: Tạo ra mục từ mới UNL – Tiếng Việt

Giao diện một mục từ trong UNL Explorer:



Vùng 1: tên mục từ Tiếng Anh

Vùng 2: Định nghĩa UNL cho mục từ

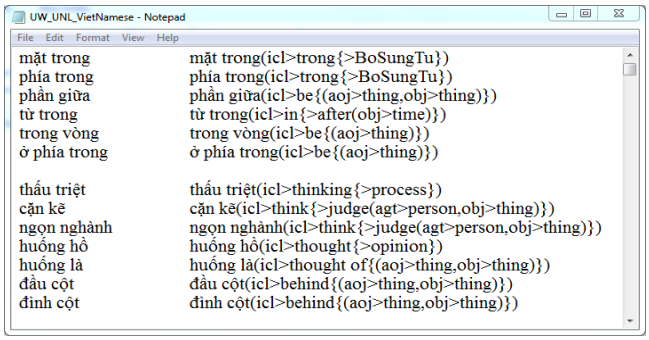
Vùng 3: Định nghĩa mục từ ngôn ngữ mới

5.5. Đánh giá

Qua quá trình triển khai, chúng tôi đã thực hiện bổ sung

thêm 30.000 mục từ (dựa trên các từ điển sẵn có) và nhập mới 550 từ (dựa trên từ điển giấy).

Mục từ UNL – Tiếng Việt lưu dưới cấu trúc sau:



Hình 4. *Cấu trúc từ điển UNL – Tiếng Việt*

Đây là một kết quả khả quan để tiếp tục việc nghiên cứu và xây dựng từ điển UNL – Tiếng Việt trở nên phong phú và chính xác hơn.

Chúng tôi xin đưa ra một số hướng phát triển để xây dựng từ điển dựa trên nguồn dữ liệu khá lớn của công cụ UNL Explorer đối với tiếng Việt như sau:

Thứ nhất: Cần nghiên cứu cấu trúc từ điển của UNL – Tiếng Anh, Anh – Việt và cấu trúc chung của UNL. Từ đó sử dụng các thành phần nhằm định nghĩa các mục từ làm tăng số lượng từ vựng trên công cụ UNL Explorer.

Thứ hai: Sử dụng một số từ điển dành cho tiếng Việt nhằm khai thác dữ liệu là từ vựng trong tất cả các lĩnh vực. Để có thể định nghĩa một số mục từ thuộc các từ loại khác nhau như: danh từ, giới từ, phó từ, tính từ,…

Thứ ba: Thường xuyên cập nhật từ điển UNL – Tiếng Việt từ UNL Explorer để tham gia phát triển hoàn chỉnh nguồn dữ liệu là từ điển UNL – Tiếng Việt. Nhằm tạo ra bộ từ điển UNL – Tiếng Việt trở nên hoàn chỉnh và lớn hơn.

6. Kết luận

Trong bài bào này, chúng tôi trình bày một quá trình bán tự động để tạo từ điển UNL – Tiếng Việt thông qua việc sử dụng các nguồn tài nguyên có sẵn trên công cụ UNL Explorer và từ điển Anh – Việt. Mặc dù đây là một quá trình không hoàn toàn tự động, nhưng đã bổ sung thêm rất nhiều số lượng mục từ vào điển UNL–Tiếng Việt.

TÀI LIỆU THAM KHẢO

[1] N. H. Siêu, L. T. Giang, and V. T. Hùng (2010), Nghiên cứu xây dựng từ điển cho hệ thống dịch tự động UNL – Tiếng Việt, Tạp chí khoa học và công nghệ, Đại học Đà Nẵng – số 4(39).

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

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

[4] W3Techs, "Usage of content languages for websites", 2011, http://www.informatik.uni-leipzig.de/~duc/Dict/W3Techs

[5] Baldwin T., Pool J., Colowick S. PanLex and LEXTRACT: Translating all Words of all Languages of the World, 2010

[6] Boguslavsky I., Cardeñosa J., Gallardo C., Iraola L. The UNL Initiative: An Overview, Computational Linguistics and Intelligent Text Processing, 2005

[7] Boguslavsky I., Dikonov V. Universal Dictionary of Concepts, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference, “Dialog 2009”

[8] Dikonov V. G. Modal Attributes in UNL, Proceedings of the 32-nd Conference “Information technologies and systems (ITiS’09)”], Bekasovo, 2009. pp. 230–237. ISBN 978-5-901158-11-1.

(BBT nhận bài: 28/07/2014, phản biện xong: 21/10/2014)