*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.2017.000118*

# XÂY DỰNG WORDNET CHO TIẾNG VIỆT

**Lâm Nhựt Khang, Nguyễn Huỳnh Hữu Đức, Võ Lê Minh Trung**

Khoa Công nghệ Thông tin và Truyền thông
Trường Đại học Cần Thơ

*lnkhang@cit.ctu.edu.vn, huuduc72216@gmail.com, votrung017@gmail.com*

***TÓM TẮT:*** *Cơ sở dữ liệu từ vựng hay mạng từ (WordNet) là nguồn tài nguyên từ vựng được sử dụng phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên như tóm tắt văn bản, rút trích thông tin và máy dịch. Bài báo trình bày phương pháp xây dựng WordNet cho tiếng Việt (VWN). Mục tiêu của nghiên cứu là xây dựng VWN có cùng cấu trúc với Princeton WordNet (PWN). Đầu tiên, các synset trong PWN được dịch về tiếng Việt để tạo ra các ứng viên tiềm năng (candidates). Một phương pháp ranking được sử dụng để loại bỏ các mục dịch không chính xác. Nhằm tăng độ bao phủ (coverage) về số lượng các synset trong VWN so với PWN, WordNet có cùng cấu trúc với PWN ở các ngôn ngữ khác nhau sẽ được sử dụng. Cuối cùng, mối quan hệ giữa các synset trong VWN được thiết lập dựa trên các mối quan hệ của synset trong PWN. VWN hiện tại chứa 78.285 synset (tỷ lệ bao phủ của synset là 66,54%) và 80.413 mối quan hệ ngữ nghĩa.*

***Từ khóa:*** *WordNet, mạng từ, synset, quan hệ ngữ nghĩa.*

## I. GIỚI THIỆU

PWN [1] (tiếng Anh) là WordNet lớn nhất trên thế giới, được xây dựng thủ công bởi các chuyên gia từ những năm 1990. Hiện tại, PWN vẫn đang trong quá trình xây dựng và hoàn thiện. WordNet là một ontology từ vựng, trong đó các từ có cùng ngữ nghĩa (cognitive synonym) được nhóm lại thành các nhóm từ đồng nghĩa. Mỗi nhóm từ đồng nghĩa được gọi là một *synset*. Các từ trong cùng một synset sẽ có ngữ nghĩa giống nhau trong một ngữ cảnh nào đó. Một từ đa nghĩa sẽ thuộc nhiều hơn một synset. Trong PWN, mỗi synset có duy nhất một *synsetid*. Có 4 loại từ loại (Part-Of-Speech –POS) trong PWN: danh từ (noun –n ), động từ (verb –v), tính từ (adjective –a) và trạng từ (adverb –r). Ví dụ: danh từ "spring" có tất cả 6 nghĩa. Bảng 1 trình bày các synset chứa danh từ "spring": synsetid của mỗi synset, thành viên trong synset đó và ngữ nghĩa (gloss) tương ứng của từng synset.

**Bảng 1.** Các synset chứa danh từ "spring" trong PWN

| Synsetid | Các từ trong cùng synset | Gloss |
|---|---|---|
| 115237044 | spring, springtime | the season of growth |
| 104288272 | spring | a metal elastic device that returns to its shape or position when pushed or pulled or pressed |
| 109443453 | spring, fountain, outflow, outpouring, natural spring | a natural flow of ground water |
| 108508361 | spring | a point at which water issues forth |
| 105021151 | give, spring, springiness | the elasticity of something that can be stretched and returns to its original length |
| 100120202 | leap, leaping, spring, saltation, bound, bounce | a light, self-propelled movement upwards or forwards |

Các synset trong WordNet liên kết với nhau bằng các mối quan hệ ngữ nghĩa (semantic relation). **Bảng *2*** trình bày một số mối quan hệ ngữ nghĩa trong PWN.

**Bảng 2.** Ví dụ một số quan hệ ngữ nghĩa trong PWN

| Quan hệ | Giải thích | Ví dụ |
|---|---|---|
| Hyponym | Y là *hyponym* của X nếu mọi Y là một loại (kind of) của X | Tập *hyponym* của từ "canid" là {bitch, dog, wolf, jackal, hyena, hyaena, fox} |
| Hypernym | Nếu Y là hyponym của X thì X là *hypernym* của Y | Tập *hypernym* của từ "dog" là {canid, canine} |
| Meronym | Y là *meronym* của X nếuY là một phần của (part of) X | Tập *meronym* của từ "chair" là {back, backrest, leg} |
| Holonym | Nếu Y là meronym của X thì X là *holonym* của Y | Từ "chair" là *holonym* của {back, backrest, leg} |
| Troponym | Y là một *troponym* của X nếu Y là một cách thức thể hiện khác của X | Từ "run" là *troponym* của từ "walk" |

Theo Vossen [2], có hai phương pháp cơ bản để xây dựng một WordNet ở ngôn ngữ đích T: phương pháp mở rộng (expand approach) và phương pháp hợp nhất (merge approach). Sử dụng phương pháp mở rộng, các chuyên gia thực hiện dịch PWN về ngôn ngữ T [3] [4] [5] và [6]. Trong phương pháp hợp nhất, WordNet ở ngôn ngữ T được xây dựng trước, sau đó các chuyên gia sẽ căn chỉnh (align) WordNet ở ngôn ngữ T với PWN [7] và [8]. Về tổng quát, WordNet được xây dựng bằng phương pháp hợp nhất sẽ có chất lượng tốt hơn về mặt ngữ nghĩa và các mối quan hệ từ vựng. WordNet được xây dựng bằng phương pháp mở rộng sẽ được thực hiện nhanh hơn và giống cấu trúc Princeton WordNet. Phương pháp mở rộng được sử dụng phổ biến hơn phương pháp hợp nhất.

Mục tiêu nghiên cứu của đề tài là xây dựng WordNet cho tiếng Việt (gọi tắt là VWN) (i) gồm các synset có độ chính xác không thấp hơn độ chính xác của các nguồn tài nguyên sẵn có được sử dụng, (ii) tỷ lệ bao phủ synset trong VWN so với PWN đạt trên 50%, (iii) thiết lập được mối quan hệ ngữ nghĩa giữa các synset và (iv) VWN sẽ có cùng cấu trúc với PWN. Các phần tiếp theo của bài báo trình bày chi tiết phương pháp xây dựng VWN. Kết quả thực nghiệm và thảo luận được trình bày trong phần IV. Cuối cùng phần V sẽ tổng kết nghiên cứu đã thực hiện.

## II. TẠO CÁC SYNSET TIẾNG VIỆT

Một trong những mục tiêu của đề tài là xây dựng VWN có cùng cấu trúc với PWN nên phương pháp mở rộng sẽ được áp dụng. Phương pháp đề xuất xây dựng VWN được phát triển dựa trên nghiên cứu của Lam et al. [6]. Nhóm tác giả xây dựng WordNet synset cho các ngôn ngữ đích T bằng cách dịch PWN về ngôn ngữ T bằng máy dịch của Microsoft[1]. Các phương pháp DR, IW và IWND được sử dụng để cải thiện chất lượng dịch. Trong phương pháp DR, các synset ở ngôn ngữ T được tạo ra bằng cách dịch trực tiếp các synset của PWN về ngôn ngữ đích. Phương pháp IW giúp làm giảm nhập nhằng ngữ nghĩa trong quá trình dịch bằng cách sử dụng WordNet ở các ngôn ngữ khác nhau có cùng cấu trúc với PWN: với mỗi synsetid trong PWN, tác giả rút trích tất cả các synset của các WordNet trung gian có cùng synsetid đó và dịch chúng về ngôn ngữ T. Đối tượng của nghiên cứu bao gồm cả ngôn ngữ nghèo tài nguyên như Karbi và Dimasa[2] nên phương pháp IWND dịch các synset có cùng synsetid ở các ngôn ngữ khác nhau về tiếng Anh, rồi tiếp tục dịch từ tiếng Anh về ngôn ngữ T. Để lựa chọn kết quả dịch có độ chính xác tốt nhất, phương pháp ranking dựa trên số lần xuất hiện (occurrence count) của ứng viên được sử dụng. Tác giả lần lượt sử dụng từ 1 đến 4 WordNet trung gian để xây dựng các WordNet mới. Thông qua thực nghiệm, họ kết luận phương pháp IW với 4 WordNet trung gian giúp xây dựng WordNet synset có chất lượng tốt nhất. Tác giả dừng lại ở giai đoạn xây dựng synset cho WordNet, chưa thiết lập các mối quan hệ ngữ nghĩa giữa các synset.

Phương pháp IW được sử dụng để xây dựng synset cho VWN. Tuy nhiên, phương pháp IW xây dựng WordNet mới có độ bao phủ synset thấp hơn phương pháp DR và IWND. Để đảm bảo vẫn tạo được synset có chất lượng và độ bao phủ synset tốt, chúng tôi cải tiến phương pháp lựa chọn ứng viên bằng cách kết hợp phương pháp ranking dựa trên số lần xuất hiện của từ [6] và phương pháp Normalized Google Distance (NGD) [9]. Bên cạnh đó, ngoài các WordNet mà Lam et al. đã thực nghiệm, chúng tôi sử dụng thêm Thai WordNet. **Bảng *3*** trình bày thông tin về các WordNet trung gian sử dụng để xây dựng VWN. Phương pháp tạo synset cho VWN được minh họa trong **Hình 1**.

**Bảng 3.** Thông tin về các WordNet trung gian sử dụng

| WordNet | Số lượng synset |
|---|---|
| PWN | 117.659 |
| FinnWordNet (FWN) [10] | 116.763 |
| Japanese WordNet (JWN) [3] | 57.184 |
| WOLF Wordnet(WWN) [7] | 59.091 |
| Thai WordNet (TWN) [11] | 73.350 |

Mỗi synset được dịch ra tiếng Việt có thể có nhiều ứng viên tiềm năng. Để lựa chọn ứng viên có độ chính xác cao, phương pháp ranking dựa trên số lần xuất hiện của từ được áp dụng để tính giá trị *rank* cho mỗi ứng viên. Rank của từ *w* (được gọi là $rank_w$) tính theo công thức:

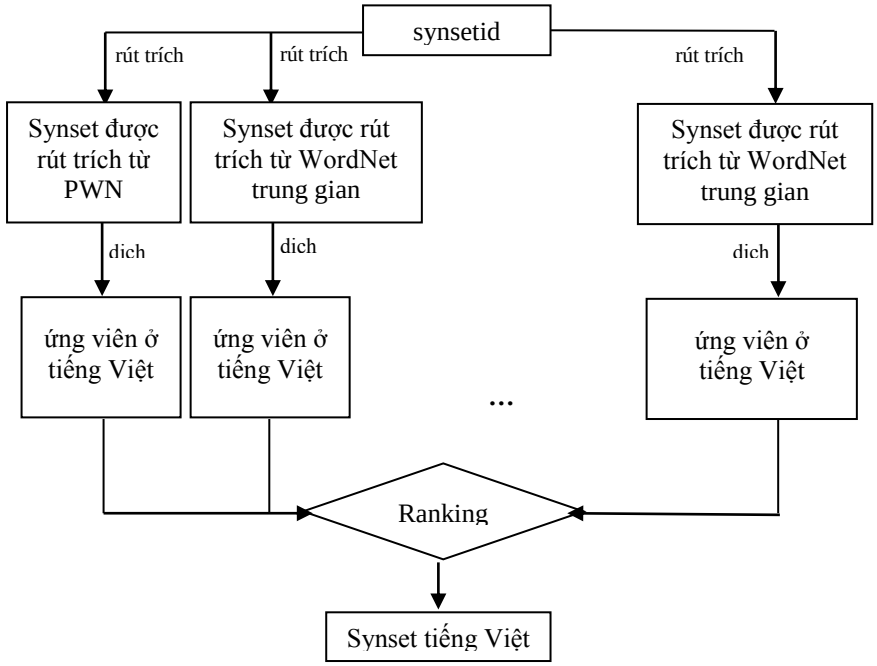
$$rank_w = \frac{occur_w}{numCandidates} * \frac{numDstWordnets}{numWordnets}$$

Trong đó:

- *numCandidates* là tổng số ứng viên của một synset
- $occur_w$ là số lần xuất hiện của từ *w* trong *numCandidates*.
- *numWordNets* là số lượng WordNet trung gian được sử dụng.
- *numDstWordNets* là số lượng WordNet trung gian chứa các từ mà các từ đó dịch sang tiếng Việt là *w*

[1] https://www.bing.com/translator

[2] Đây là các ngôn ngữ "nguy cấp", gần như chỉ có một từ điển duy nhất giữa nó và tiếng Anh. Các máy dịch hiện tại không hỗ trợ các ngôn ngữ này.




**Hình 1.** Tạo synset cho VWN

Giá trị rank nằm trong khoảng từ 0,00 đến 1,00. Một ứng viên với giá trị rank càng cao thì khả năng trở thành từ thích hợp thuộc synset càng lớn. Việc lựa chọn ứng viên được thực hiện như sau:

### *A. Trường hợp 1:*

Tất cả các ứng viên có chỉ số rank l,00 được chấp nhận như là một thành viên của synset đó trong VWN. Điều này có nghĩa là tất cả các từ trong synset đó ở tất cả các WordNet trung gian đều được dịch về cùng một từ trong tiếng Việt. Số lượng WordNet trung gian được sử dụng càng nhiều thì mức độ dịch chính xác của ứng viên với rank 1,00 càng lớn.

**Bảng *4*** trình bày các thành viên của synset có synsetid 100001740 với gloss là "that which is perceived or known or inferred to have its own distinct existence (living or nonliving)" được rút trích từ các WordNet khác nhau. Do tất cả các thành viên trong synset ở các WordNet khác nhau đều dịch về tiếng Việt có nghĩa là "thực thể" và rank của từ "thực thể" là 1,00 nên "thực thể" được xem như là thành viên của synset có synsetid 100001740.

**Bảng 4.** Ví dụ trường hợp 1 – ứng viên có rank 1,00

| WordNet | Thành viên trong synset | Ứng viên tiềm năng | Rank |
|---|---|---|---|
| PWN | entity | thực thể | 1,00 |
| WWN | entité | thực thể | |
| JPN | 実体 | thực thể | |

### *B. Trường hợp 2:*

Nếu synset không có ứng viên nào có rank là 1,00 thì tất cả các ứng viên có rank cao nhất đều được chấp nhận là thành viên của synset đó. Ví dụ, **Bảng *5*** trình bày các thành viên của của synset có synsetid 200010435 với gloss "behave in a certain manner; show a certain behavior; conduct or comport oneself" được rút trích từ các WordNet trung gian tương ứng. Do không có ứng viên nào có rank 1,00 và ứng viên "cư xử" có rank cao nhất nên "cư xử" được chấp nhận là một thành viên của synset có synsetid 200010435 trong VWN.

**Bảng 5.** Ví dụ trường hợp 2- chọn ứng viên có rank cao nhất (khác 1,00)

| WordNet | Thành viên trong synset | Ứng viên tiềm năng | Rank |
|---|---|---|---|
| PWN | act | hành động | 0,213 |
| | behave | cư xử | 0,400 |
| | do | làm | 0,053 |
| FWN | toimia | hành động | 0,213 |
| | tehdä | cư xử | 0,400 |
| | käyttäytyä | làm | 0,053 |
| WWN | agir | hành động | 0,213 |

| WordNet | Thành viên trong synset | Ứng viên tiềm năng | Rank |
|---|---|---|---|
| | se comporter | cư xử | 0,400 |
| JWN | 振る舞う | cư xử | 0,400 |
| | 振舞う | cư xử | 0,400 |
| | 立ち振る舞う | đứng lên và hành động | 0,013 |
| | 行動+する | hành động | 0,213 |
| TWN | ประพฤติตัว | cư xử | 0,400 |
| | มีพฤติกรรม | thái độ | 0,027 |
| | มีความประพฤติ | thái độ | 0,027 |

***C. Trường hợp 3:***

Nếu tất cả các ứng viên trong cùng synset có cùng chỉ số rank cao nhất (khác 1,00), phương pháp NGD được sử dụng để lựa chọn ứng viên thích hợp. Khoảng cách NGD giữa từ $w_1$ và $w_2$ được tính theo công thức:

$$\mathrm{NGD}(w_1, w_2) = \frac{\max\{\log f(w_1), \log f(w_2)\} - \log f(w_1, w_2)}{\log M - \min\{\log f(w_1), \log f(w_2)\}}$$

Trong đó:

- $M$ là tổng số lượng các trang web được được tìm kiếm bởi Google. Hiện tại, giá trị của M khoảng 50.500.000.000[3].
- $f(w_1)$ và $f(w_2)$ lần lượt là số lượng kết quả tìm kiếm được của $w_1$ và $w_2$.
- $f(w_1,w_2)$ là số lượng kết quả tìm kiếm được chứa đồng thời $w_1$ và $w_2$.

Giá trị NGD từ 0 đến ∞. Giá trị này càng gần về 0 thì nghĩa của cặp từ tìm kiếm càng giống nhau. Do đó, nếu tồn tại 2 ứng viên có khoảng cách NGD nhỏ hơn ngưỡng α, chúng ta sẽ chấp nhận chúng như là thành viên của synset. Giá trị ngưỡng α được chọn dựa trên thực nghiệm. Chúng tôi tính toán giá trị NGD cho các cặp ứng viên tiềm năng trong 50 synset ngẫu nhiên, giá trị α thay đổi 1,00; 0,75; 0,50; 0,40; 0,35; 0,32 và 0,20. Sau đó thực hiện đánh giá độ tương đồng về ngữ nghĩa của các cặp từ. Giá trị α giúp xác định các từ tương đồng ngữ nghĩa tốt và độ bao phủ cao là 0,32, nên đây cũng là giá trị được chọn trong quá trình thực nghiệm xây dựng toàn bộ VWN.

Ví dụ, xét synset có synsetid 110399491 với gloss "a father or mother; one who begets or one who gives birth to or nurtures and raises a child; a relative who plays the role of guardian" gồm các thành viên và kết quả dịch tương ứng của ứng viên sang tiếng Việt được trình bày trong **Bảng *6***.

**Bảng 6.** Thành viên của synset có synsetid 110399491

| WordNet | Thành viên trong synset | Ứng viên tiềm năng | Rank |
|---|---|---|---|
| PWN | parent | cha mẹ | 0,083 |
| FWN | äiti | mẹ | 0,083 |
| | isä | cha | 0,083 |
| JWN | ペアレント | phụ huynh | 0,083 |

Tất cả ứng viên tiềm năng đều có rank bằng nhau và khác 1,00, nên phương pháp NGD được sử dụng để tính khoảng cách giữa các ứng viên tiềm năng. Ví dụ, tổng số trang web indexed hiện tại của Google là 50.500.000.000, số lượng kết quả tìm được của "cha mẹ", "phụ huynh" và "cha mẹ, phụ huynh" theo thứ tự lần lượt là 655.000, 515.000 và 20.700. Vậy NGD (cha mẹ, phụ huynh) là 0,300. Thực hiện tương tự, giá trị NGD của các cặp ứng viên tiềm năng còn lại theo thứ tự (cha mẹ, mẹ), (cha mẹ, cha), (mẹ, cha), (mẹ, phụ huynh) và (cha, phụ huynh) là 0,567; 0,568; 1,198; 0,773 và 0,936. Ngưỡng α là 0,32, phương pháp NGD sẽ loại bỏ ứng viên "cha" và "mẹ", chỉ chấp nhận "cha mẹ" và "phụ huynh" là thành viên của synset có synsetid 110399491 trong VWN.

## III. THIẾT LẬP CÁC MỐI QUAN HỆ NGỮ NGHĨA TRONG VWN

Sau khi tạo được các synset cho VWN, bước kế tiếp là thiết lập mối quan hệ ngữ nghĩa giữa các synset, được trình bày trong Giải thuật 1. Đầu tiên, từng $synset_{Vi}$ trong VWN sẽ được ánh xạ (map) với $synset_{Pj}$ tương ứng trong PWN thông qua giá trị synsetid (Giải thuật 1, dòng 1-2). Với mỗi $synset_{Pj}$, thực hiện trích xuất tất cả các mối quan hệ ngữ nghĩa $sem\_relation_r$ và các $synset_{Pk}$ tương ứng trong mối quan hệ đó (Giải thuật 1, dòng 3-4). Kế tiếp, kiểm tra sự

[3] http://www.worldwidewebsize.com/

tồn tại của $synset_{Vu}$ trong VWN tương ứng với $synset_{Pk}$ trong PWN (Giải thuật 1, dòng 5-6), nếu $synset_{Vu}$ tồn tại trong VWN, thêm mối quan hệ $sem\_relation_r$ tương ứng giữa $synset_{Vi}$ và $synset_{Vu}$ vào VWN (Giải thuật 1, dòng 7-8). Tên các mối quan hệ ngữ nghĩa giữa các synset trong PWN được giữ nguyên ở tiếng Anh, không dịch về tiếng Việt.

**Giải thuật 1**: Thiết lập mối quan hệ ngữ nghĩa giữa các synset trong VWN

```
1: For all synset_Vi in VWN created
2:     synset_Pj ← map(synset_Vi, PWN)
3:     For all synset_Pj in PWN
4:         Extract all sem_relation_r (synset_Pj, synset_Pk)
5:         For all sem_relation_r (synset_Pj, synset_Pk)
6:             synset_Vvu ← map(synset_Pk, VWN)
7:             If exist synset_Vu then
8:                 add sem_relation_r (synset_Vi, synset_Vu)
9:             end if
10:        end for
11:    end for
12: end for
```

Tiếp theo ví dụ ở mục II (trường hợp 3), trong PWN, synsetid 110399491 gồm một thành viên {parent}. Sau khi dịch về tiếng Việt thì synset này gồm 2 thành viên là {phụ huynh, cha mẹ}. **Bảng *7*** trình bày một số mối quan hệ ngữ nghĩa giữa synset này và các synset khác trong PWN và VWN sau khi áp dụng Giải thuật 1.

**Bảng 7.** *Một số mối quan hệ ngữ nghĩa của synset có synsetid 110399491*

| Synset 1 | Synset 2 | | | | Mối quan hệ ngữ nghĩa |
|---|---|---|---|---|---|
| | Synsetid | Thành viên trong synset | | Gloss | |
| | | PWN | VWN | | |
| 110399491 | 107970406 | {family, family unit} | {gia đình, hộ gia đình} | "primary social group; parents and children" | member meronym |
| 110399491 | 109772448 | {adopter, adoptive parent} | {cha mẹ nuôi} | "a person who adopts a child of other parents as his or her own child" | hyponym |
| 110399491 | 110332385 | {female parent, mother } | {mẹ } | "a woman who has given birth to a child (also used as a term of address to your mother" | hyponym |
| 110399491 | 110126708 | {genitor} | {cha mẹ ruột} | "a natural father or mother" | hypernym |
| 110399491 | 110654932 | {stepparent} | {cha dượng} | "the spouse of your parent by a subsequent marriage" | hyponym |
| 110399491 | 109918248 | {kid, child} | {đứa trẻ} | "a human offspring (son or daughter) of any age" | antonym |

Một điều cần lưu ý là mối quan hệ ngữ nghĩa giữa $synset_1$ và $synset_2$ có thể khác với mối quan hệ ngữ nghĩa của $synset_2$ đối với $synset_1$. Nói cách khác $sem\_relation(synsetid_1, synsetid_2)$ có thể không giống với $sem\_relation(synsetid_2, synsetid_1)$. Ví dụ, trong **Bảng *7*** sem_relation(110399491, 110654932) là hyponym; nhưng chiều ngược lại, sem_relation(110654932, 110399491) là hypernym. Tuy nhiên, sem_relation(110399491, 109918248) và sem_relation(109918248, 110399491) đều là antonym.

## IV. KẾT QUẢ THỰC NGHIỆM

### *A. Tài nguyên sử dụng*

VWN được xây dựng dựa trên PWN phiên bản 3.0. PWN 3.0 có tổng cộng 117.659 synset, trong đó gồm 82.115 synset danh từ, 13.767 synset động từ, 18.156 synset tính từ và 3.621 synset trạng từ. Các WordNet trung gian khác được sử dụng là FWN, WWN, JWN và TWN với các thông tin như đã trình bày ở Bảng 3.

Chất lượng synset của VWN phụ thuộc rất lớn vào API dịch từ các ngôn ngữ trung gian về tiếng Việt. Lam et al [6] sử dụng Microsoft Translator API để dịch các WordNet về ngôn ngữ đích. Vào thời điểm thực nghiệm, Microsoft Translator API không cho phép sử dụng miễn phí nữa, nên Yandex Translate API[4] là một lựa chọn thay thế.

Sau khi xây dựng các synset và thiết lập các mối quan hệ của synset, chúng tôi sẽ quản lý VWN dựa theo dự án WNSQL[5].

### *B. Phương pháp đánh giá*

Phương pháp tiêu chuẩn để đánh giá một WordNet mới xây dựng ở ngôn ngữ T là yêu cầu người dùng đánh giá toàn bộ nguồn tài nguyên đó. Người đánh giá là các chuyên gia ngôn ngữ T hoặc ít nhất phải sử dụng thành thạo ngôn ngữ T. VWN vừa được xây dựng đang được từng bước chỉnh sửa và cải tiến chất lượng. Do đó, để đánh giá VWN bước đầu xây dựng được, 500 synset và các quan hệ ngữ nghĩa của chúng trong VWN được chọn ngẫu nhiên để đánh giá. 8 người dùng sử dụng ngôn ngữ tiếng Việt như tiếng mẹ đẻ được yêu cầu đánh giá các sysnet này sử dụng thang 5-điểm: 5: chính xác (Excellent), 4: tốt (Good), 3: trung bình (Average), 2: tạm chấp nhận (Fair) và 1: sai (Bad).

### *C. Kết quả*

Các VWN lần lượt được xây dựng bằng phương pháp chúng tôi đề xuất (ký hiệu IW-NGD) và phương pháp IW [6] với số lượng WordNet trung gian lần lượt là 4 (PWN, FWN, WWN và JPN) và 5 (PWN, FWN, WWN, JPN và TWN). **Bảng *8*** trình bày số lượng synset xây dựng được, độ bao phủ synset so với PWN và điểm trung bình đánh giá của các synset trong các VWN. Kết quả thực nghiệm cho thấy VWN được xây dựng bằng 5 WordNet trung gian có chất lượng và độ bao phủ tốt hơn sử dụng 4 WordNet trung gian. Bên cạnh đó, VWN xây dựng bằng phương pháp chúng tôi đề xuất IW-NGD có chất lượng và độ bao phủ tốt hơn VWN được xây dựng bằng phương pháp IW.

**Bảng 8.** Số lượng synset và điểm trung bình đánh giá của synset

| Phương pháp | Số WordNet trung gian | Số lượng synset | Điểm đánh giá | Độ bao phủ |
|---|---|---|---|---|
| IW | 4 | 55.048 | 3,21 | 46,79% |
| IW | 5 | 61.808 | 3,61 | 52,53% |
| IW-NGD | 4 | 61.348 | 3,23 | 52,14% |
| IW-NGD | 5 | 78.285 | 3,73 | 66,54% |

VWN được xây dựng bằng phương pháp IW-NGD với 5 WordNet trung gian đạt độ chính xác và độ bao phủ synset tốt nhất trong thực nghiệm; do đó, các synset trong VWN này được thiết lập mối quan hệ ngữ nghĩa. Tổng số 80.413 mối quan hệ ngữ nghĩa giữa các synset/từ được thiết lập và điểm đánh giá trung bình của các mối quan hệ này là 3,70.

### *D. Thảo luận*

VWN được xây dựng bởi Lam et al. [6] sử dụng phương pháp IW với 4 WordNet trung gian có 72.010 synset (độ bao phủ 61.20%) và điểm trung bình đánh giá là 4.26/5.00. VWN tạo được của Lam et al. sử dụng phương pháp IW và 4 WordNet trung gian có kết quả tốt hơn VWN của chúng tôi cũng được xây dựng hoàn toàn dựa trên phương pháp của họ (IW và 4 WordNet trung gian), chỉ khác duy nhất là họ sử dụng Microsoft Translator API và chúng tôi sử dụng Yandex Translate API. Điều này giúp chúng tôi đưa ra kết luận là chất lượng dịch của Yandex API thấp hơn so với Microsoft Translator API. Thêm vào đó, với mỗi từ ở ngôn ngữ nguồn, phần lớn các API hỗ trợ dịch chỉ trả về một kết quả dịch duy nhất hoặc không dịch được từ tương đương ở tiếng Việt. Vì vậy, kết quả dịch đạt độ chính xác không cao, đặc biệt là trong trường hợp các từ đa nghĩa hoặc từ có nhiều loại từ. Ví dụ: từ "book" có nghĩa là "quyển sách" trong trường hợp danh từ hoặc sẽ có nghĩa là "đặt chỗ" trong trường hợp động từ, tuy nhiên API hỗ trợ dịch chỉ trả về một nghĩa duy nhất "quyển sách". Để VWN hoàn hiện hơn, các API dịch sẽ được thay thế bằng các từ điển song ngữ có chất lượng.

VWN được quản lý dựa theo dự án WNSQL. Trong WNSQL định nghĩa 14 bảng dữ liệu: adjpositions, adjpositiontypes, casedwords, lexdomains, linktypes, lexlinks, semlinks, morph, morphmaps, samples, senses, synsets, words, postypes. Trong PWN, tổng cộng có 28 loại mối quan hệ ngữ nghĩa liệt kê trong bảng linktypes. Thực tế, chỉ có 26 loại mối quan hệ giữa các synset được hình thành, mối quan hệ "domain" và "member" không tìm thấy. Đối với VWN, chỉ có 16 loại mối quan hệ được thiết lập. Một trong những lý do là do synset trong VWN chưa bao phủ hết các synset trong PWN. Để cải thiện điều này, trước tiên chất lượng dịch synset từ PWN về VWN cần được cải thiện. Bên cạnh đó, do sự khác biệt về ngôn ngữ và văn hóa nên các mối quan hệ của synset trong PWN có thể không giống với các mối quan hệ của synset trong VWN. Để VWN hoàn thiện hơn, nghiên cứu sẽ rất cần sự hỗ trợ giúp đỡ của các nhà ngôn ngữ học.

[4] https://tech.yandex.com/translate/
[5] http://wnsql.sourceforge.net/

Đề tài Mạng từ tiếng Việt[6] (gọi tắt là MTTV) của Bộ Khoa học và Công nghệ xây dựng mạng từ tiếng Việt dựa trên 3 nhóm loại từ: danh từ, động từ và tính từ. Đề tài được chia làm 2 giai đoạn [12] (i) dịch phần lõi của WordNet tiếng Anh ra tiếng Việt, phần lõi được chọn là các từ có tần suất xuất hiện cao trong kho dữ liệu tiếng Anh BNC[7] (ii) bổ sung các khái niệm chỉ có ở tiếng Việt vào mạng từ. Hiện tại, MTTV đang có 40.788 synset và 67.344 từ. Đối với VWN, dữ liệu từ vựng trong các WordNet ở các ngôn ngữ khác nhau có cùng cấu trúc với PWN được khai thác tối đa, điều này giúp cho VWN có độ bao phủ synset tốt nhất có thể, không phải chỉ bao gồm phần lõi của PWN. VWN được xây dựng cho cả 4 nhóm từ loại: danh từ, động từ, tính từ và cả trạng từ. Việc nghiên cứu đối sánh VWN và MTTV là cần thiết; và để giúp xây dựng một WordNet hoàn chỉnh hơn cho tiếng Việt, một trong các bước tiếp theo của đề tài là tích hợp dữ liệu của MTTV và VWN.

## V. KẾT LUẬN

Mục đích của nghiên cứu là bước đầu xây dựng WordNet cho tiếng Việt có cùng cấu trúc với PWN và độ bao phủ synset tốt nhất có thể. Các synset và các mối quan hệ giữa chúng trong VWN đã được xây dựng. Bước kế tiếp, chất lượng synset và độ bao phủ của synset trong VWN sẽ được cải thiện bằng cách sử dụng các từ điển song ngữ (thay thế cho API dịch) kết hợp với vectơ biểu diễn từ (vector representation of words) để tìm hiểu mối quan hệ giữa các từ trong văn bản (sử dụng kho ngữ liệu Wikipedia[8]), giúp cải thiện mối quan hệ giữa các từ trong cùng synset và cả các mối quan hệ ngữ nghĩa trong việc xây dựng WordNet mới. Chúng tôi đang triển khai xây dựng gloss cho các synset, đồng thời kêu gọi sự hỗ trợ của các chuyên gia giúp đánh giá và hiệu chỉnh toàn diện các synset, mối quan hệ ngữ nghĩa đã xây dựng.

## TÀI LIỆU THAM KHẢO


[1] C. Fellbaum, WordNet, Blackwell Publishing Ltd, 1998.

[2] P. Vossen, "Building WordNets," 2005.

[3] Hiroyuki Kaji and Mariko Watanabe, "Automatic construction of Japanese WordNet," in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006.

[4] Mortaza Montazery and Heshaam Faili, "Automatic Persian WordNet construction," in *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, 2010.

[5] Martin Saveski and Igor Trajkovski, "Automatic construction of WordNets by using machine translation and language modeling," in *Proceedings of the 13th International MultiConference Information Society, volume C*, Ljubljana, Slovenia, 2010.

[6] Lam, Khang N., Al Tarouti, F. and Kalita, J., "Automatically constructing Wordnet Synsets," in *ACL*, 2014.

[7] Benoit Sagot and Darja Fiser, "Building a free French WordNet from multilingual resources," in *Ontolex*, Marrakech, Morocco, 2008.

[8] Debasri Chakrabarti, Vaijayanthi Sarma, and Pushpak Bhattacharyya, "Complex predicates in Indian language WordNets," *Lexical Resources and Evaluation Journal*, Vols. 40(3-4), 2007.

[9] Cilibrasi, Rudi and Paul Vitanyi, "Automatic meaning discovery using Google," in *Dagstuhl Seminar Proceedings. Schloss Dagstuhl-Leibniz-Zentrum für Informatik*, 2006.

[10] Lindén, Krister, and Lauri Carlson, "FinnWordNet–Finnish WordNet by Translation," *LexicoNordica–Nordic Journal of Lexicography*, vol. 17, pp. 119-140, 2010.

[11] Thoongsup S., Charoenporn T., Robkop K., Sinthurahat T., Mokarat C., Sornlertlamvanich V., Isahara H, "Thai Wordnet Construction," in *The 7th Workshop on Asian Language Resources (ALR7), Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP)*, Suntec, Singapore, 2009.

[12] Nguyễn Phương Thái, Phạm Văn Lam, Nguyễn Hoàng Trung, Trần Ngọc Anh, Trương Thị Thu Hà, "Cộng đồng xử lý tiếng Việt," 28 3 2015. [Online]. Available: http://wordnet.vn/vi/chi-tiet/tong-quan-ve-xay-dung-mang-tu-tieng-viet-18-1.html. [Accessed 2017].


[6] http://viet.wordnet.vn/wnms/

[7] http://www.natcorp.ox.ac.uk/

[8] https://www.wikipedia.org/

# CONSTRUCTING A VIETNAMESE WORDNET

**Lam Nhut Khang, Nguyen Huynh Huu Duc, Vo Le Minh Trung**

***ABSTRACT:*** *A lexical database or WordNet is an important resource and widely used in many research fields such as natural language processing, document summarizaton, information retrieval or machine translation. This study presents the begining step to construct a Vietnamese WordNet (VWN). One of our goals is to build a VWN having the same structure as the Princeton WordNet (PWN). Therefore, our approach translates synsets in PWN to Vietnamese using a translation API to generate translation candiates. Then, a ranking method is applied on these candidates to find the best translations. To increase the coverage percentage of synsets in VWN compared to the PWN, WordNets linked to the PWN in other languages are used as intermediate resources. Finally, semantic relations in VWN are created by extracting semantic relations of corresponding synsets in PWN. Our VWN consists of 78,285 synsets with the coverage percentage of 66.54% and 80,413 semantic relations.*