*Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Nha Trang, ngày 8-9/10/2020*
*DOI: 10.15625/vap.2020.00158*

# GÁN NHÃN NGỮ NGHĨA CHO TỪ ĐIỂN TIẾNG VIỆT DỰA TRÊN WORDNET TIẾNG VIỆT

**Lê Phạm Ngọc Yến, Trần Thị Anh Thư, Đinh Điền**
Trường Đại học Khoa học Tự nhiên, ĐHQG TP. HCM
*lpnyen@apcs.vn, ttathu@apcs.vn, ddien@fit.hcmus.edu.vn*

***TÓM TẮT:*** *Từ điển là một trong những kho tàng tri thức quý báu của mỗi ngôn ngữ bởi nó bao hàm hầu hết các từ vựng và định nghĩa tương ứng của mỗi từ vựng trong ngôn ngữ đó. Chính vì vậy, từ điển có thể trở thành nguồn ngữ liệu quan trọng cho các tác vụ khai thác văn bản (text mining), dịch máy (machine translation) hay phân tích cảm xúc đối tượng (sentiment analysis). Trong đó, một thách thức được đặt ra là mỗi từ vựng trong từ điển cần phải được đánh dấu theo một số những tiêu chí phân loại hữu ích. Ngoài việc gán nhãn từ loại (part-of-speech tagging), gán nhãn ngữ nghĩa (semantic tagging) sẽ giúp đánh dấu các từ vựng trong từ điển dựa trên mặt ngữ nghĩa của từ, phục vụ đắc lực cho việc ứng dụng tri thức vào xử lý ngôn ngữ tự nhiên. Trong báo cáo này, nhóm tác giả đề xuất phương pháp gán nhãn ngữ nghĩa cho từ điển tiếng Việt bằng cách ánh xạ giữa từ vựng trong từ điển và từ nguyên mẫu (lemma) trong tập đồng nghĩa (synset) của WordNet tiếng Việt. Việc đánh dấu sẽ được thể hiện trong cấu trúc dữ liệu của từ điển bởi một trường dữ liệu ghi nhận những định danh của những tập đồng nghĩa (synset id) tương ứng với từ vựng đó.*

***Từ khóa:*** *Gán nhãn ngữ nghĩa, từ điển, WordNet tiếng Việt.*

## I. GIỚI THIỆU

Từ điển tiếng Việt là công trình khoa học nghiên cứu về tiếng Việt một cách khách quan và phổ quát. Trong đó, hệ thống từ vựng và giải nghĩa từ tiếng Việt được tổ chức theo những quy tắc và cấu trúc chặt chẽ [1]. Vì vậy, từ điển tiếng Việt vốn được dùng như một công cụ chuẩn mực cho việc tra cứu tiếng Việt, mang tính định hướng quan trọng đối với quá trình học tập và vận dụng ngôn ngữ tiếng Việt cho người học. Với sự xuất hiện của ngành xử lý ngôn ngữ tự nhiên (một nhánh quan trọng của lĩnh vực Trí tuệ Nhân tạo), nhu cầu trợ giúp máy móc hiểu và xử lý ngôn ngữ tự nhiên nói chung và tiếng Việt nói riêng càng trở nên thiết yếu. Chính nhờ những đặc thù về cấu trúc dữ liệu và giá trị ngôn ngữ nói trên mà từ điển tiếng Việt có thể trở thành nguồn ngữ liệu hữu ích cho các tác vụ xử lý ngôn ngữ tự nhiên.

Ở từ điển tiếng Việt, nhãn từ loại thường là loại thông tin thông dụng vì loại thông tin này đã được xây dựng bài bản và có thể truy vấn trực tiếp. Tuy nhiên, thế mạnh của từ điển lại nằm chủ yếu ở phần thông tin về ngữ nghĩa của từ hay nói cách khác là phần giải nghĩa từ. Loại thông tin này lại chưa được lượng hóa theo một dạng phù hợp cho công tác tính toán. Vì vậy, khi làm việc với mặt nghĩa của từ, bản thể luận (ontology) thường được sử dụng thay thế. Một trong những mô hình bản thể luận ngôn ngữ đa tri thức có tầm ảnh hưởng lớn là hệ WordNet [2]. WordNet là một từ điển ý niệm nổi tiếng được xây dựng bởi các nhà khoa học tại đại học Princeton. WordNet nguyên bản được nghiên cứu và xây dựng dựa trên ngữ nghĩa của từ vựng tiếng Anh. Các từ vựng (từ nguyên mẫu hay lemma) trong WordNet được tổ chức thành các tập đồng nghĩa (synset) dưới dạng cây và mạng lưới thể hiện mối quan hệ giữa các tập. Do cách tổ chức này mà WordNet được sử dụng nhiều trong việc gán nhãn dữ liệu, phân tích ngữ nghĩa từ, dịch máy,… Hiện nay, nhiều ngôn ngữ đã tiến hành phát triển hệ WordNet cho riêng mình, trong đó có tiếng Việt. Do một số hệ WordNet tiếng Việt được xây dựng tự động dựa trên WordNet nguyên bản của tiếng Anh [3], việc đảm bảo độ che phủ giữa hai ngôn ngữ trở nên một thách thức lớn: WordNet tiếng Việt chưa thể bao hàm hết các từ vựng trong tiếng Việt. Trên thực tế, nhóm nghiên cứu đã kiểm nghiệm được tỉ lệ từ vựng trong từ điển tiếng Việt có mặt (về dạng từ) trong WordNet tiếng Việt chỉ đạt xấp xỉ 43,09 %. Điều này cho thấy việc sử dụng trực tiếp bản thể luận WordNet tiếng Việt cho các tác vụ tính toán ngôn ngữ tiếng Việt sẽ khó cho hiệu quả cao. Nhu cầu thiết yếu đặt ra là làm thế nào để tận dụng được kho từ vựng đa dạng và phổ quát của tiếng Việt ở từ điển trong khi lại có thể lượng hóa được ngữ nghĩa của chúng trong tính toán?

Nhằm góp phần giải quyết vấn đề này, nhóm nghiên cứu tiến hành đề xuất giải pháp gán nhãn ngữ nghĩa tự động cho từ điển tiếng Việt dựa trên WordNet tiếng Việt. Về cơ bản, giải pháp này nhằm giúp tăng cường cho từ điển tiếng Việt thông tin thêm về nghĩa của từ vựng. Thông tin này đóng vai trò là con trỏ chỉ đến tập đồng nghĩa trong WordNet tiếng Việt có chứa hay có liên quan đến từ vựng đó. Với độ bao phủ đạt xấp xỉ 95,44 %, giải pháp này sẽ phần nào giúp từ điển hoàn thành vai trò của một kho ngữ liệu tính toán được về cả dạng từ lẫn nghĩa của từ.

Bài báo này được tổ chức thành 4 mục. Mục 2 trình bày chi tiết về giải pháp mà nhóm tác giả đề xuất. Mục 3 báo cáo kết quả của giải pháp. Mục 4 đưa ra kết luận cho nghiên cứu, đề xuất một số hướng cải thiện và phát triển cho giải pháp.

## II. NGHIÊN CỨU THỰC NGHIỆM

### *A. Tổng quan dữ liệu*

*1.* Từ điển tiếng Việt - VDic

Từ điển VDic (thuộc bản quyền của Trung tâm Ngôn ngữ học Tính toán, Trường ĐH Khoa học Tự nhiên - ĐHQG TP. HCM) gồm 7 trường dữ liệu ghi nhận thông tin về mặt hình thái từ và nghĩa của từ, trong đó 3 trường dữ liệu được quan tâm trong bài nghiên cứu gồm có: Từ vựng (Word), Nhãn từ loại (POS) và Giải nghĩa-Ví dụ

(Explanation-Examples). Về dung lượng, VDic có 42414 khái niệm, bao gồm: nhóm biểu tượng thông dụng (dấu câu, dấu phép toán,…), tên khoa học/đơn vị đo lường quốc tế dạng nguyên bản và Việt hóa (*Ag* - nguyên tố hóa học, *alpha, kilogram*,…), từ viết tắt nước ngoài (*AUD* - Đô la nước Úc,…), từ nước ngoài thông dụng (*picnic, javel*,…) và 41510 từ gồm các danh từ, tính từ, động từ, trạng từ, thán từ,… thuần Việt, từ phiên âm, cặp từ hô ứng, thành ngữ và tên viết tắt tiếng Việt. Phần Giải nghĩa - ví dụ gồm hai loại: 1) một danh sách gồm các giải nghĩa của từ và câu ví dụ có chứa từ vựng đó trong ngữ cảnh của nghĩa tương ứng; 2) con trỏ chỉ tới từ đồng nghĩa. Do hiện tượng từ đồng âm và từ nhiều nghĩa của tiếng Việt mà một dạng từ có thể xuất hiện nhiều lần tùy theo số lượng nhãn từ loại có thể có của nó. Vì vậy, việc gán nhãn ngữ nghĩa cũng sẽ phải phù hợp với điều này.

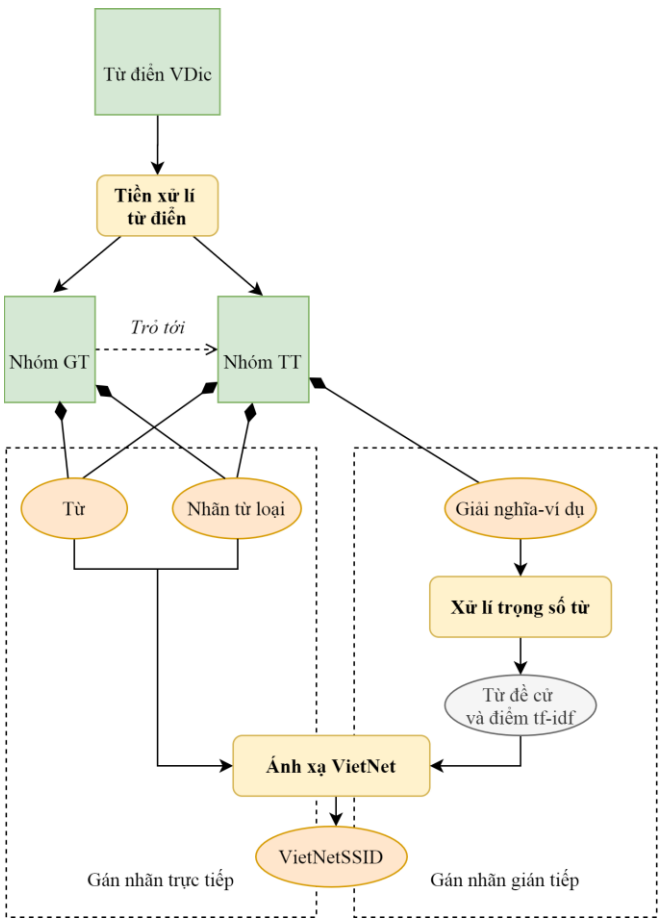
*2.* Hệ WordNet tiếng Việt - VietNet

VietNet [3] (thuộc bản quyền của Trung tâm Ngôn ngữ học Tính toán, Trường ĐH Khoa học Tự nhiên - ĐHQG TP. HCM) có số lượng tập đồng nghĩa và số lượng từ tương đồng với WordNet, trong đó con số cụ thể là 117658 tập đồng nghĩa, 82115 danh từ, 13766 động từ, 36312 tính từ và 3621 trạng từ. Mỗi tập đồng nghĩa có một định danh (synset id) riêng biệt. Các mối quan hệ ngữ nghĩa giữa các từ trong VietNet cũng được thể hiện tương tự như WordNet, trong đó bao gồm: quan hệ đồng nghĩa (synonymy), quan hệ trái nghĩa (antonymy), quan hệ thuộc cấp (hyponymy), quan hệ bao hàm (hypernymy), quan hệ bộ phận (meronymy), quan hệ tổng thể (holonymy), quan hệ kéo theo (entailment) và quan hệ cách thức đặc biệt (troponymy).

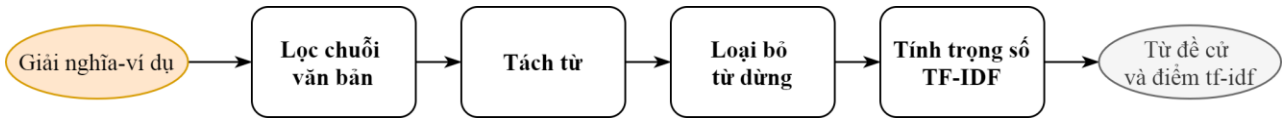
Quan hệ đồng nghĩa thể hiện tính tương đồng giữa các nghĩa của từ. Đây cũng là quan hệ quan trọng nhất trong VietNet vì cấu trúc tổ chức dữ liệu được hình thành dựa vào mối quan hệ này là chính. Giữa hai từ thì chỉ số đánh giá sự đồng nghĩa được thuộc khoảng 0 đến 1; nếu là 0 thì hoàn toàn không đồng nghĩa (không nhất thiết là trái nghĩa) và nếu là 1 thì là đồng nghĩa hoàn toàn. Trái với quan hệ đồng nghĩa là quan hệ trái nghĩa. Cũng như trong quan hệ đồng nghĩa, hai từ có mối quan hệ trái nghĩa không nhất thiết là phủ định của nhau.

Quan hệ thuộc cấp và quan hệ bao hàm là quan hệ ngữ nghĩa giữa các nghĩa của từ (word meanings). Ví dụ, *bộ_bò_biển* là thuộc cấp của *bộ_động_vật*. Quan hệ bộ phận chỉ rằng từ này là một bộ phận của từ khác. Ví dụ, *giống_lợn_biển* là bộ phận của *bộ_bò_biển*. Quan hệ kéo theo được dùng cho động từ, thể hiện ý niệm rằng nếu động từ 1 được thực thi thì động từ 2 cũng được thực thi. Ví dụ, *ngáy* kéo theo *ngủ*. Quan hệ cách thức đặc biệt chỉ rằng động từ 1 là một hình thức thực hiện động từ 2 thông qua cách thức cụ thể nào đó. Ví dụ, *thì thầm* là cách thức đặc biệt của *nói*.

### *B. Phương pháp nghiên cứu*




Hình 1. Tổng quan quy trình gán nhãn ngữ nghĩa cho từ điển tiếng Việt vDic




Hình 2. Các bước tìm Từ đề cử và tính điểm TF-IDF

Mục tiêu của việc gán nhãn ngữ nghĩa cho từ điển tiếng Việt VDic là với mỗi từ vựng $w$ trong từ điển, có thể tìm được một hay một số những tập đồng nghĩa trong VietNet có liên quan về ngữ nghĩa với từ vựng đó. Như vậy, vấn đề đặt ra là sự liên quan về ngữ nghĩa này cần được xác định như thế nào. Do mỗi từ vựng trong từ điển được giải thích bằng sự kết hợp của những từ vựng khác, nên việc xác định xem một từ vựng có liên quan đến những khái niệm nào thì phần giải nghĩa từ cần phải được đưa vào xem xét. Vì vậy, nhóm tác giả đề xuất hai tiêu chí đánh giá như sau: 1) dựa trên dạng từ và nhãn từ loại của Từ vựng; 2) dựa trên dạng từ và điểm TF-IDF của các Từ đề cử thuộc phần Giải nghĩa - ví dụ của từ vựng đó. Tổng quan của toàn bộ quá trình được trình bày ở Hình 1, trong đó quy trình gán nhãn trực tiếp được tiến hành dựa trên tiêu chí 1 và quy trình gán nhãn gián tiếp được tiến hành dựa trên tiêu chí 2. Kết quả của hai quá trình gán nhãn này được gọi là VietNetSSID, là tập hợp các định danh của tập đồng nghĩa kèm theo trọng số của chúng.

Ở VDic, vì phần Giải nghĩa - ví dụ gồm hai loại là giải nghĩa trực tiếp và con trỏ đến từ đồng nghĩa, nên sau khi tiền xử lý từ điển về kiểu dữ liệu của các trường, kho từ điển lớn này được chia thành hai phần: nhóm từ được giải nghĩa trực tiếp (Nhóm TT) và nhóm từ được giải nghĩa gián tiếp (Nhóm GT) tương ứng với hai loại nêu trên. Sau khi các từ thuộc Nhóm TT được gán nhãn ngữ nghĩa, dữ liệu của các từ thuộc Nhóm GT sẽ được cập nhật theo con trỏ của chúng.

*1.* Gán nhãn trực tiếp

VietNet chỉ bao gồm 4 nhóm từ loại chính là danh từ, động từ, tính từ và trạng từ trong khi VDic lại chia từ loại một cách chi tiết hơn như danh từ riêng, danh từ chỉ loại, danh từ đơn vị, cảm từ, tính từ tượng thanh, tính từ tượng hình, định từ, từ viết tắt, thành ngữ,... nên trước khi truy vấn VietNet, hai tập từ loại này được quy định sao cho tương ứng với nhau. Thao tác truy vấn định danh của tập đồng nghĩa trong VietNet diễn ra đơn giản với đầu vào là Từ vựng và Nhãn từ loại, kết quả đầu ra là chuỗi định danh. Trọng số của định danh của tập đồng nghĩa tìm được trong quá trình này được gán là 1.0/1.0.

*2.* Gán nhãn gián tiếp

Chi tiết phần xử lý Từ đề cử của quy trình gán nhãn gián tiếp được trình bày ở Hình 2.

**Lọc chuỗi văn bản:**

Trong trường Giải nghĩa - ví dụ của từ điển được đánh dấu bởi các kí hiệu phân chia nội dung này thành nhiều phần: phần giải nghĩa, phần cho ví dụ, các cờ đánh dấu đặc điểm từ (từ thông tục, từ cũ,...), một số chú thích về cách dùng, hàm ý,... Ngoài ra, một số nghĩa của các từ nhiều nghĩa còn có thêm dạng con trỏ đến nghĩa khác. Như vậy, nhóm tác giả đã sử dụng biểu thức chính quy để lọc ra được hai loại chuỗi văn bản cần thiết gồm chuỗi văn bản giải nghĩa và chuỗi văn bản ví dụ.

**Tách từ:**

Phương pháp được sử dụng để chia ranh giới giữa các từ trong câu mà nhóm nghiên cứu sử dụng là phương pháp truy vấn từ điển. Vì số lượng từ vựng có độ dài trên 3 chữ chỉ chiếm 5,99 % kho từ vựng và đa số là thành ngữ đã được tạo nên bởi các từ có nghĩa nên độ dài tối đa khi cắt từ được chọn là 3. Phương pháp này được trình bày đơn giản như sau: với đầu vào là một chuỗi gồm $N$ chữ, lần lượt xem xét từng cụm tối đa $n$ chữ ($n = 3$). Nếu cụm này tồn tại trong từ điển (ở đây, sử dụng từ điển VDic) thì lượt xem xét tiếp theo bắt đầu từ chữ kế tiếp của chữ cuối cùng trong cụm vừa cắt ở trong câu. Trường hợp hiếm gặp là sau khi dò cụm một chữ mà cụm vẫn không có mặt trong từ điển, cụm này vẫn được công nhận là một từ độc lập.

**Loại bỏ từ dừng:**

Dựa vào danh sách từ dừng của tiếng Việt, ngay sau khi một cụm từ được tách thành công, nếu nó có mặt trong danh sách từ dừng thì sẽ bị loại bỏ khỏi câu.

**Tính trọng số TF-IDF:**

TF-IDF là một phương pháp thống kê trong xử lý ngôn ngữ tự nhiên, thể hiện chỉ số quan trọng của từ trong văn bản. Trong bài báo này, mỗi chuỗi văn bản đầu vào nêu trên được xem như một "văn bản". Vì thế, điểm TF-IDF của mỗi từ đối với một văn bản sẽ thể hiện độ quan trọng của nó đối với một định nghĩa trong từ điển. Hay nói cách khác, điểm TF-IDF của một từ trong một định nghĩa thể hiện sự đóng góp của nó trong việc xác định từ vựng tương ứng. Công thức TF-IDF được áp dụng trong quá trình này như sau:

- $\text{tf-idf} = \text{tf}(t, d) \times \text{idf}(t)$
- $\text{tf} = \frac{f_{t,d}}{\sum_{t' \in d}(f_{t',d})}$

trong đó $f_{t,d}$ là tần suất xuất hiện của từ $t$ trong định nghĩa $d$.

- $\text{idf}(t) = \log\left(\frac{1+n}{1+\text{df}(t)}\right) + 1$

trong đó $n$ tổng số định nghĩa trong từ điển và $\text{df}(t)$ là tổng số định nghĩa có chứa từ $t$.

Như vậy, kết quả đầu ra của giai đoạn này bao gồm Từ đề cử liên quan đến Từ vựng tương ứng của chúng về mặt nghĩa, kèm theo đó là điểm số đánh giá được mức độ liên quan, chúng được sắp xếp giảm dần theo giá trị của điểm số TF-IDF. Kết quả này được tiếp tục mang vào xử lý ở bước Ánh xạ VietNet. Cách truy vấn tương tự như quy trình gán nhãn trực tiếp nêu trên nhưng không sử dụng thông tin về từ loại. Mỗi định danh của tập đồng nghĩa trả về được gán trọng số TF-IDF tương ứng với Từ đề cử của nó. Vì không sử dụng thông tin từ loại nên đây là một điểm gây nhiễu cho kết quả cuối cùng mà nhóm tác giả nhận thấy cần phải cải thiện.

## III. KẾT QUẢ

Trong quá trình gán nhãn trực tiếp, khi không sử dụng thông tin Giải nghĩa - ví dụ của Từ vựng, số từ được gán nhãn ngữ nghĩa trong VDic chỉ đạt 43,09 %. Lý do cho việc này nằm ở sự khác biệt về dạng từ của từ vựng giữa hai bộ dữ liệu: từ vựng trong từ điển VDic gần gũi với tiếng Việt và đời sống trong khi từ vựng của VietNet mang tính học thuật và gồm nhiều từ là kết quả của quá trình khái quát hóa khái niệm trong mạng lưới các tập đồng nghĩa như *dạng_người*, *phần_thân*,... Một số từ trong từ điển không thể được gán nhãn qua quy trình này có thể kể đến như *gọt dũa*, *ham hố*, *heo hắt*, *hiền thục*, *hiển vinh*,... dù cho những từ này đều thông dụng trong tiếng Việt.

Khi kết hợp với quá trình gán nhãn gián tiếp có sử dụng thông tin Giải nghĩa - ví dụ của Từ vựng kèm trọng số, tổng số từ được gán nhãn ngữ nghĩa lên đến 40480 từ, đạt xấp xỉ 95,44 %. Một kết quả ngẫu nhiên được chọn báo cáo là từ vựng *gieo rắc* được trình bày trong Bảng 1. Trong đó, ô VietNetSSID gồm định danh của tập đồng nghĩa và trọng số tương ứng là kết quả tổng hợp sau 2 quy trình. Tập đồng nghĩa trả về cho thấy không chỉ có những nét nghĩa trực tiếp được sử dụng gán nhãn cho từ gieo rắc mà có thêm những nét nghĩa mở rộng được ghi nhận như: mô tả về thao tác (*ném*, *phóng*, *tung*, *quẳng*), hình thái sự vật, sự việc (*rơi*, *ngã*, *chạm*, *lan truyền*), hàm ý (*tai họa*, *tai hại*). Bên cạnh đó, một số tập đồng nghĩa trả về lại không liên quan đến từ *gieo rắc* (hồ tiêu, đầu tư).

| **Từ vựng:** *gieo rắc* | **Từ loại:** động từ | **Giải nghĩa-ví dụ:** @* (id.) Làm cho rơi xuống khắp nơi trên một diện rộng, gây hậu quả tai hại@• Ném bom gieo rắc chất độc hoá học@* Đưa đến và làm cho lan truyền rộng (thường là cái xấu, cái tiêu cực)@• Gieo rắc hoang mang. Chiến tranh gieo rắc đau thương tang tóc |
|---|---|---|
| **Từ đề cử và trọng số:** [('rắc', 0.5387), ('gieo', 0.5023), ('đau_thương', 0.2066), ('lan_truyền', 0.2022), ('tai_hại', 0.194), ('chất_độc', 0.1852), ('bom', 0.1745), ('hậu_quả', 0.1735), ('tang', 0.1719), ('ném', 0.1683), ('diện', 0.1651), ('hoang', 0.1637), ('chiến_tranh', 0.157), ('hoá_học', 0.1536), ('rơi', 0.1521), ('tiêu', 0.1453), ('khắp', 0.1404), ('rộng', 0.1316)] | | |
| **VietNetSSID:** | | |
| [('01477417-v', 1), ('01019354-v', 1), ('00945290-v', 1), ('01477980-v', 0.5387), ('08246176-n', 0.5387), ('00433149-n', 0.5387), ('02063061-v', 0.5387), ('01590484-v', 0.5387), ('01515964-v', 0.5387), ('01353412-v', 0.5387), ('01351212-v', 0.5387), ('01350014-v', 0.5387), ('01348064-v', 0.5387), ('00548903-v', 0.5387), ('00329914-v', 0.5387), ('01544361-v', 0.5023), ('01477081-v', 0.5023), ('00716079-v', 0.5023), ('08370270-n', 0.2066), ('02066095-v', 0.2022), ('02060385-v', 0.2022), ('01813190-v', 0.2022), ('01355160-v', 0.2022), ('01033921-v', 0.2022), ('00945766-v', 0.2022), ('00944427-v', 0.2022), ('00914305-v', 0.2022), ('00263505-v', 0.2022), ('00471241-n', 0.194), ('00060492-a', 0.194), ('02359925-a', 0.194), ('01076146-a', 0.194), ('00973734-a', 0.194), ('00865369-a', 0.194), ('00208180-a', 0.194), ('00298767-r', 0.194), ('16760832-n', 0.1852), ('04058422-n', 0.1745), ('03250907-n', 0.1745), ('12792201-n', 0.1735), ('12791219-n', 0.1735), ('08150346-n', 0.1735), ('05798039-n', 0.1735), ('15583803-n', 0.1719), ('15363192-n', 0.1719), ('04109366-n', 0.1719), ('01469491-n', 0.1683), ('00127110-n', 0.1683), ('00126085-n', 0.1683), ('00125209-n', 0.1683), ('00123457-n', 0.1683), ('02202016-v', 0.1683), ('01874697-v', 0.1683), ('01839095-v', 0.1683), ('01584971-v', 0.1683), ('01573803-v', 0.1683), ('01567968-v', 0.1683), ('01491651-v', 0.1683), ('01488417-v', 0.1683), ('01488268-v', 0.1683), ('01485593-v', 0.1683), ('01485495-v', 0.1683), ('01485096-v', 0.1683), ('01484354-v', 0.1683), ('01483130-v', 0.1683), ('01411630-v', 0.1683), ('01372298-v', 0.1683), ('01371401-v', 0.1683), ('01213844-v', 0.1683), ('01057160-v', 0.1683), ('16162471-n', 0.1651), ('02606760-v', 0.1651), ('02123809-v', 0.1651), ('00043391-v', 0.1651), ('00042397-v', 0.1651), ('00786813-a', 0.1651), ('02235645-a', 0.1637), ('00163707-r', 0.1637), ('01135927-n', 0.157), ('01124680-n', 0.157), ('01110959-n', 0.157), ('06777032-n', 0.1536), ('02520242-a', 0.1536), ('02519899-a', 0.1536), ('08227158-n', 0.1521), ('02731220-v', 0.1521), ('02586979-v', 0.1521), ('01959730-v', 0.1521), ('01953997-v', 0.1521), ('01951874-v', 0.1521), ('01947100-v', 0.1521), ('01540575-v', 0.1521), ('01518978-v', 0.1521), ('00143446-v', 0.1521), ('02331953-a', 0.1521), ('14998099-n', 0.1453), ('14637032-n', 0.1453), ('14636763-n', 0.1453), ('13791371-n', 0.1453), ('08730849-n', 0.1453), ('08110443-n', 0.1453), ('04456571-n', 0.1453), ('03417171-n', 0.1453), ('01313488-n', 0.1453), ('02347489-v', 0.1453), ('02281310-v', 0.1453), ('02246485-v', 0.1453), ('02008582-v', 0.1453), ('00472366-v', 0.1453), ('00215565-r', 0.1404), ('00187164-r', 0.1404), ('02394146-a', 0.1316), ('01344206-a', 0.1316), ('02394886-a', | | |

0.1316), ('02394537-a', 0.1316), ('01740498-a', 0.1316), ('01333668-a', 0.1316), ('01282975-a', 0.1316), ('01282648-a', 0.1316), ('01033006-a', 0.1316), ('00487335-a', 0.1316), ('00487041-a', 0.1316), ('00485791-a', 0.1316), ('00099004-a', 0.1316), ('00375585-r', 0.1316)]

**Tên synset tương ứng:**

[('vãi.v.07'): 1, ('gieo_rắc.v.02'): 1, ('gieo_rắc.v.03'): 1, ('rải.v.04'): 0.5387, ('rắc.n.01'): 0.5387, ('rắc.n.02'): 0.5387, ('bỏ_rải_rác.v.01'): 0.5387, ('rắc.v.02'): 0.5387, ('rắc.v.03'): 0.5387, ('rắc.v.05'): 0.5387, ('bỏ_rải_rác.v.03'): 0.5387, ('rắc.v.07'): 0.5387, ('rắc.v.08'): 0.5387, ('lắc.v.11'): 0.5387, ('nứt_nẻ.v.01'): 0.5387, ('trồng.v.03'): 0.5023, ('gieo.v.04'): 0.5023, ('in_sâu.v.01'): 0.5023, ('quẫn.n.02'): 0.2066, ('lan_truyền.v.01'): 0.2022, ('truyền.v.09'): 0.2022, ('lan_truyền.v.03'): 0.2022, ('lan_truyền.v.04'): 0.2022, ('truyền.v.16'): 0.2022, ('truyền_bá.v.03'): 0.2022, ('phổ_biến.v.10'): 0.2022, ('lộ_ra.v.03'): 0.2022, ('truyền_bá.v.07'): 0.2022, ('phương_hại.n.01'): 0.194, ('bất_lợi.a.02'): 0.194, ('tai_hại.s.02'): 0.194, ('có_hại.s.05'): 0.194, ('tai_họa.s.01'): 0.194, ('tai_hại.s.05'): 0.194, ('tai_hại.s.06'): 0.194, ('tổn_hại.r.01'): 0.194, ('chất_độc.n.01'): 0.1852, ('bom.n.01'): 0.1745, ('bom.n.02'): 0.1745, ('hậu_quả.n.01'): 0.1735, ('ảnh_hưởng.n.09'): 0.1735, ('hậu_quả.n.03'): 0.1735, ('hậu_quả.n.04'): 0.1735, ('sự_đau_buồn.n.01'): 0.1719, ('tiếp_tuyến.n.02'): 0.1719, ('trống_định_âm.n.01'): 0.1719, ('ném.n.01'): 0.1683, ('ném.n.02'): 0.1683, ('phóng.n.13'): 0.1683, ('quẳng.n.02'): 0.1683, ('ném.n.05'): 0.1683, ('phóng.v.01'): 0.1683, ('phóng.v.09'): 0.1683, ('chuyển_bất_ngờ.v.01'): 0.1683, ('tống.v.02'): 0.1683, ('ném.v.05'): 0.1683, ('ném.v.06'): 0.1683, ('ném.v.07'): 0.1683, ('dồi.v.03'): 0.1683, ('ném.v.09'): 0.1683, ('bạt.v.02'): 0.1683, ('phóng.v.13'): 0.1683, ('ném.v.12'): 0.1683, ('đẩy.v.13'): 0.1683, ('phóng.v.14'): 0.1683, ('tung.v.07'): 0.1683, ('giã.v.02'): 0.1683, ('đánh_đập.v.02'): 0.1683, ('ném_phịch.v.01'): 0.1683, ('sút.v.02'): 0.1683, ('phạm_vi.n.01'): 0.1651, ('trưng_diện.v.01'): 0.1651, ('phô_trương.v.03'): 0.1651, ('phục_sức.v.02'): 0.1651, ('chải_chuốt.v.06'): 0.1651, ('chải_chuốt.s.03'): 0.1651, ('dại.a.01'): 0.1637, ('dại.r.01'): 0.1637, ('binh_đao.n.01'): 0.157, ('chiến_sự.n.03'): 0.157, ('cuộc_đấu.n.06'): 0.157, ('hóa_học.n.02'): 0.1536, ('hóa_học.a.01'): 0.1536, ('hóa_học.a.02'): 0.1536, ('rơi.n.01'): 0.1521, ('rơi.v.01'): 0.1521, ('chạm.v.01'): 0.1521, ('ngã.v.02'): 0.1521, ('đậu.v.04'): 0.1521, ('rơi.v.05'): 0.1521, ('rơi.v.06'): 0.1521, ('hạ_cánh_khẩn_cấp.v.02'): 0.1521, ('nhỏ.v.03'): 0.1521, ('rơi.v.09'): 0.1521, ('rơi.s.01'): 0.1521, ('đi_tiêu.n.01'): 0.1453, ('hồ_tiêu.n.01'): 0.1453, ('tiêu.n.03'): 0.1453, ('chuối.n.01'): 0.1453, ('tiêu.n.05'): 0.1453, ('cột_mốc.n.02'): 0.1453, ('sáo_dọc.n.01'): 0.1453, ('ống_tiêu.n.03'): 0.1453, ('tiêu.n.09'): 0.1453, ('đầu_tư.v.01'): 0.1453, ('chi.v.01'): 0.1453, ('tiêu.v.03'): 0.1453, ('làm_phân_tán.v.01'): 0.1453, ('tiêu.v.05'): 0.1453, ('khắp.r.01'): 0.1404, ('chi_chít.r.01'): 0.1404, ('rộng_rãi.a.01'): 0.1316, ('rộng.a.02'): 0.1316, ('sâu.s.03'): 0.1316, ('rộng.s.04'): 0.1316, ('rộng_rãi.s.05'): 0.1316, ('rộng.s.06'): 0.1316, ('rộng.s.07'): 0.1316, ('rộng.s.08'): 0.1316, ('rộng_bụng.s.02'): 0.1316, ('rộng.s.10'): 0.1316, ('rộng_rãi.s.09'): 0.1316, ('rộng_rãi.s.10'): 0.1316, ('phồng.s.02'): 0.1316, ('rộng.r.01'): 0.1316]

Bằng khảo sát thủ công trên 50 từ được gán nhãn, nhóm tác giả thấy được việc gán nhãn trực tiếp đôi khi lại hoàn toàn sai lệch về nghĩa giữa từ vựng và tập đồng nghĩa tìm được. Lý do cho việc này nằm ở quá trình xây dựng VietNet tự động và hiện tượng từ đồng nghĩa, từ nhiều nghĩa trong tiếng Việt. Tuy nhiên, khi sử dụng những Từ để cử của Từ vựng để gán nhãn, bằng tri thức con người, nhóm tác giả thấy được sự xuất hiện của những tập đồng nghĩa liên quan về nghĩa cho Từ vựng đó. Đối với vấn đề này, nhóm nghiên cứu chưa đề xuất được phương pháp đánh giá cho độ chính xác về nghĩa giữa nghĩa của Từ vựng trong từ điển VDic và ý niệm của tập đồng nghĩa tương ứng trong VietNet.

## IV. KẾT LUẬN

Việc gán nhãn ngữ nghĩa cho từ điển tiếng Việt giúp tăng cường ý nghĩa cho các từ vựng. Kết quả nghiên cứu cho thấy phần giải nghĩa của từ vựng sau khi gán nhãn đã được mở rộng và liên kết với những nét nghĩa mới. Những nét nghĩa này lại là những ý niệm được tổ chức chặt chẽ trong hệ bản thể luận VietNet, vì vậy các thao tác tính toán về ngữ nghĩa sẽ trở nên khả thi và có cơ sở khoa học. Điều này giúp từ điển tiếng Việt trở thành nguồn ngữ liệu đa năng, phục vụ được nhu cầu sử dụng cả dạng từ và ngữ nghĩa của từ. Một điểm hạn chế trong giải pháp của nhóm nghiên cứu đó là chưa đề xuất được phương pháp đánh giá độ chính xác của quá trình gán nhãn. Trong giai đoạn tiếp theo, nhóm tác giả sẽ tiến hành nghiên cứu trên vấn đề này. Ngoài ra, nhóm tác giả nhận thấy kết quả gán nhãn cần được xử lý thêm để cô đọng và chọn lọc ra những nét nghĩa tiêu biểu hơn bằng cách sử dụng những lợi thế giá trị hơn của VietNet như các mối quan hệ về ngữ nghĩa và các hệ đo tương đồng.

## V. LỜI CẢM ƠN

*Nhóm nghiên cứu xin được cảm ơn Trung tâm Ngôn ngữ học Tính toán và Chương trình Advance Program in Computer Science (APCS), thuộc Trường ĐH Khoa học Tự nhiên - ĐHQG TP. HCM đã hỗ trợ kinh phí để chúng tôi thực hiện nghiên cứu này. Cách đặc biệt, nhóm nghiên cứu sinh xin được gửi lời tri ân đến PGS.TS. Đinh Điền đã hướng dẫn đề tài và xin cảm ơn các anh chị nghiên cứu sinh của Trung tâm Ngôn ngữ học Tính toán đã chia sẻ, góp ý và hỗ trợ nhóm nghiên cứu trong suốt quá trình thực hiện đề tài.*

## TÀI LIỆU THAM KHẢO

[1] Vũ Xuân Lương, Nguyễn Thị Minh Huyền, "Building a Vietnamese Computational Lexicon", Proceedings of the National Symposium on Research, Development and Application of Information and Communication Technology, Vietnam, 2008.

[2] George A. Miller, "WordNet: A Lexical Database for English", Communications of the ACM Vol. 38, No. 11: 39-41, 1995.

[3] Trần Văn Tri, "Dịch tự động WordNet từ tiếng Anh sang tiếng Việt dựa vào từ điển Oxford Anh-Việt", Luận văn thạc sĩ, Đại học Khoa học Tự nhiên - Đại học Quốc gia TP. HCM, 2017.

# SEMANTIC TAGGING FOR VIETNAMESE DICTIONARY USING VIETNAMESE WORDNET

**Le Pham Ngoc Yen, Tran Thi Anh Thu, Dinh Dien**

***ABSTRACT:*** *Since dictionaries contain vocabularies and their definitions of a language, they can become a viable linguistic corpus for many topics in natural language processing (NLP) such as text mining, machine translation, and sentiment analysis. One of the challenges in using dictionaries for these activities is that their vocabularies need to be tagged according to some useful criteria. In addition to part-of-speech tagging, semantic tagging can help categorize words by their meanings, which is important for the utilization of semantic information in NLP tasks. In this paper, we propose an approach towards dictionary tagging by projecting words of the dictionary to lemmas in synonym sets (synsets) of the Vietnamese WordNet. In the Vietnamese dictionary data structure, the projection is presented as a field containing a list of the corresponding Vietnamese WordNet's synset identifications.*

***Keywords:*** *Semantic tagging, dictionary, Vietnamese WordNet.*