

SO SÁNH CẤU TRÚC MẠNG TỪ TIẾNG VIỆT (VIETNET) VỚI MẠNG TỪ TIẾNG ANH (WORDNET) TRÊN CÁC DANH TỪ THỰC THỂ (ENTITY)

PHAN THI MỸ TRANG¹DƯƠNG THI AN²ĐINH ĐIỀN³TRẦN THI MINH PHƯƠNG⁴

Abstract: WordNet is a well-known conceptual system (a type of a synonym dictionary, or a thesaurus) designed for the English language, which has been widely recognized and applied in research works related to computational semantics, lexical semantics, contrastive semantics, etc. VietNet is a similar conceptual system for the Vietnamese language, developed by the Computational Linguistics Center (CLC). To evaluate the structure of VietNet, we need to compare the structure of VietNet with WordNet by aligning the measures in VietNet and WordNet. In this study, the Resnik semantic similarity measure was used to compute semantic distances for 1400 pairs of Vietnamese entity nouns and 1400 pairs of English entity nouns. The comparison results showed that the semantic distance values between Vietnamese and English entity nouns are relatively close, with a small difference of 0.26. This indicates that the structure of VietNet is reliable for entity nouns, as it exhibits a high similarity with the structure of WordNet. The findings of this research are significant and can be applied in various linguistic domains, such as English-Vietnamese comparison in terms of lexical semantics and cross-lingual text similarity. Additionally, the research results highlight the potential of VietNet and the Resnik measure in comparing the semantic distance values of nouns across languages. We hope that this study will contribute to the expansion of knowledge regarding the measurement of semantic similarity between languages and serve as a foundation for further studies in this field.

Keywords: *semantic distance, WordNet, VietNet, entity noun, similarity measure*

1. Giới thiệu

Trong từ điển thông thường, các từ được sắp xếp theo thứ tự abc mà không quan tâm đến ngữ nghĩa của chúng. Để giải quyết vấn đề này, các dạng từ điển đồng nghĩa (thesaurus) ra đời, tập trung vào cung cấp các từ đồng nghĩa (synonyms) dựa trên quan hệ đồng nghĩa hoặc gần nghĩa và tổ chức phân cấp (hierarchy) hay còn gọi là cây phả hệ (ontology).

Trong tiếng Anh, có nhiều công trình về hệ thống ý niệm ngữ nghĩa như LLOCE, LDOCE, CoreLex và WordNet [1]. Trong số này, WordNet là mạng từ tiếng Anh nổi tiếng nhất, được xây dựng bởi Đại học Princeton (1986). WordNet tổ chức các từ thành các tập đồng nghĩa

¹ ThS, Trường Đại học Công nghệ Sài Gòn; (NCS Trường Đại học Khoa học Xã hội và Nhân văn, Đại học Quốc gia Thành phố Hồ Chí Minh)

² Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh

³ PGS.TS, Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh

⁴ TS, Trường Đại học Khoa học Xã hội và Nhân văn, Đại học Quốc gia Thành phố Hồ Chí Minh

(synsets) biểu diễn các ý niệm hoặc khái niệm cụ thể, và các từ (gồm danh từ, động từ, tính từ, trạng từ) được liên kết với nhau thông qua các quan hệ ngữ nghĩa. Các mối quan hệ ngữ nghĩa giữa các từ trong WordNet 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) [8].

Trong tiếng Việt, chúng tôi sử dụng mạng từ VietNet được xây dựng bởi Trung tâm Ngôn ngữ học Tính toán (CLC). Đây là một hệ thống ý niệm tiếng Việt, nhằm phục vụ cho việc phân tích về ngữ nghĩa học từ vựng, ngữ nghĩa học đối chiếu cho tiếng Việt. Hệ thống ý niệm VietNet được xây dựng dựa trên nguyên tắc tương tự như WordNet. Nhưng để đánh giá chính xác mức độ tương đồng về mặt cấu trúc giữa VietNet và WordNet, chúng tôi sử dụng độ đo tương đồng ngữ nghĩa Resnik để tính toán khoảng cách ngữ nghĩa cho các cặp danh từ thực thể trong mạng từ tiếng Việt và mạng từ tiếng Anh. Khái niệm “thực thể” (entity) trong bài báo này được hiểu như trong mạng từ tiếng Anh (WordNet): *[entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))]*⁵. “Thực thể” ở đây là một đối tượng hoặc vật thể được nhận biết, biết đến hoặc suy luận có sự tồn tại riêng biệt của nó. Thực thể bao gồm vật có sự sống (living thing, organism) và vật thể không có sự sống (non-living thing, object). Theo nguyên lý tổ chức của mạng từ tiếng Việt cũng như tiếng Anh những từ càng gần nghĩa thì độ đo khoảng cách càng nhỏ. Chúng tôi áp dụng nguyên lý này trên cả hai mạng từ tiếng Việt và tiếng Anh, sau đó quan sát sự chênh lệch về độ đo này. Nếu sự chênh lệch càng nhỏ thì sự tương đồng về cấu trúc của hai mạng từ này càng cao.

Bằng cách thu thập ngữ liệu từ VietNet và WordNet, chúng tôi đã tính toán khoảng cách ngữ nghĩa cho 1.400 cặp danh từ thực thể tiếng Việt và 1.400 cặp danh từ thực thể tiếng Anh. Kết quả so sánh cho thấy có sự tương đồng đáng kể trong cấu trúc giữa VietNet và WordNet trên các danh từ thực thể. Điều này chứng minh rằng cấu trúc của cây ý niệm VietNet này trên các danh từ thực thể có độ tin cậy cao. Kết quả này sẽ góp phần làm phong phú thêm các nghiên cứu về tương đồng ngữ nghĩa từ vựng giữa các ngôn ngữ và tạo nền tảng cho những nghiên cứu tiếp theo trong lĩnh vực này.

2. Lịch sử nghiên cứu

Đến nay, nhiều nước đã nghiên cứu xây dựng mạng từ tương tự như vậy cho ngôn ngữ của họ như mạng từ EuroWordNet cho các ngôn ngữ châu Âu, HowNet cho tiếng Hoa hay một số mạng từ cho tiếng Việt trong đó có VietNet của Trung tâm ngôn ngữ học tính toán (CLC).

Năm 1996, EuroWordNet bắt đầu dự án mạng từ cho khối Liên minh Châu Âu (EU), với mục tiêu phát triển các mạng từ cho tiếng Hà Lan, Tây Ban Nha và Ý; đồng thời liên kết các mạng từ này với mạng từ tiếng Anh trong một cơ sở dữ liệu đa ngôn ngữ. Năm 1997, dự án được mở rộng sang tiếng Đức, tiếng Pháp, tiếng Séc và tiếng Estonia. EuroWordNet được hoàn thành vào cuối năm 1999. Cơ sở dữ liệu EuroWordNet tương thích cốt lõi với Mạng từ tiếng Anh, đã được bổ sung và mở rộng đáng kể về các quan hệ và đối tượng [13].

⁵ <https://wordnet.princeton.edu>

Năm 2016, Guo và cộng sự đã đề xuất một độ đo mới để đo độ tương đồng ngữ nghĩa giữa các từ dựa trên cây ý niệm được xây dựng từ mạng từ tiếng Hoa (HowNet). Độ tương đồng giữa hai ý niệm được tính bằng công thức xem xét độ sâu của từ cha (hypernym) chung gần nhất, độ dài đường đi ngắn nhất giữa các ý niệm và các thông số điều chỉnh. Kết quả nghiên cứu cho thấy độ đo đề xuất đạt được mức độ đo tương đồng ngữ nghĩa tốt hơn so với các độ đo khác dựa trên HowNet. Tuy nhiên, nó vẫn còn một số hạn chế như không chính xác trong việc đo độ tương đồng giữa một số cặp từ cụ thể. Các tác giả đề xuất việc mở rộng và cập nhật liên tục cơ sở tri thức thế giới và cải thiện kiến trúc từ điển để giải quyết vấn đề này [5].

Trong các công trình nghiên cứu tiếng Việt, năm 2003, Duc H.N. và cộng sự đề xuất xây dựng mạng từ cho ngôn ngữ tiếng Việt bằng cách sử dụng các nguồn tài nguyên từ điển có sẵn, bao gồm Mạng từ tiếng Anh (Princeton WordNet), từ điển song ngữ Việt - Anh và từ điển đơn ngữ tiếng Việt. Nhóm tác giả đã phân tích các nguồn tài nguyên có sẵn, đánh giá khả năng áp dụng của phương pháp tự phát hiện và thiết kế công cụ tính toán để giúp việc thu thập và lọc thông tin từ vựng và ngữ nghĩa từ các từ điển MRD (Machine Readable Dictionary: từ điển có thể đọc được bằng máy tính) và các nguồn tài nguyên khác [3]. Năm 2016, Nguyễn T. P. và cộng sự trình bày các kết quả mới nhất của quá trình xây dựng dự án mạng từ cho ngôn ngữ tiếng Việt. Nhóm tác giả đề xuất sử dụng các nguồn tài nguyên ngôn ngữ có sẵn và đảm bảo các đặc điểm về ngôn ngữ và văn hóa đặc thù của tiếng Việt. Nhóm tác giả cũng đưa ra các kết quả thống kê và phân tích để chỉ ra các đặc điểm của Mạng từ tiếng Anh (Princeton WordNet) [10]. Năm 2017, nghiên cứu của Dương T. H. và cộng sự tập trung vào việc cải tiến phương pháp xây dựng thư viện từ điển cộng tác (collaborative) cho mạng từ tiếng Việt. Họ sử dụng nguyên tắc đồng thuận giữa các thành viên trong nhóm cộng tác để lựa chọn kết quả dịch tối ưu, tạo ra phiên bản đầu của mạng từ tiếng Việt. Để hoàn thiện mạng từ tiếng Việt, nhóm nghiên cứu đã phát triển ứng dụng trực tuyến của cây phả hệ có tên là Ontology Wiki (OntWiki). OntWiki cho phép người dùng tải lên và hiển thị chi tiết của mạng từ tiếng Việt. Người dùng xem được các lớp, cấu trúc của các thuộc tính đối tượng, thuộc tính dữ liệu, thuộc tính chủ thích và danh sách các cá thể theo lớp. Người dùng có thể thảo luận và đóng góp ý kiến cá nhân thông qua tài khoản của họ trên OntWiki. Kết quả nghiên cứu cho thấy phương pháp xây dựng thư viện từ điển cộng tác đã làm tăng tính hiệu quả và chất lượng đầu ra trong xây dựng thư viện từ điển tiếng Việt [4].

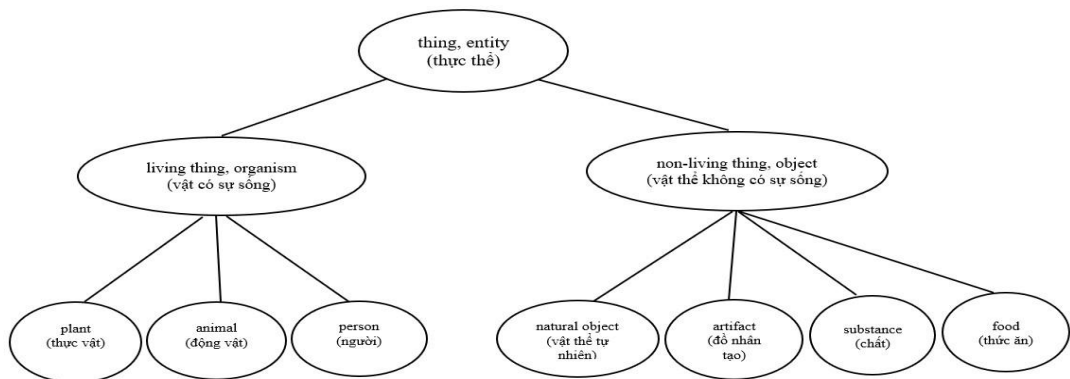
Năm 2021, Trương Hoàng Bảo và cộng sự đã xây dựng WordNet tiếng Việt tự động bằng ngữ liệu song ngữ. Trong nghiên cứu này, nhóm tác giả đã thực hiện một số cải tiến như sử dụng kho ngữ liệu TEDTalk, cải tiến phương pháp chọn lọc các tập đồng nghĩa (synsets) và đặc biệt là chỉ dựa trên WordNet của 4 ngôn ngữ nguồn là Anh, Tây Ban Nha, Pháp và Nhật Bản thay vì phải sử dụng WordNet của 14 ngôn ngữ khác nhau. Kết quả, phương pháp xây dựng của nhóm tác giả có độ chính xác (Precision) là 61,42% và WordNet tiếng Việt xây dựng được với 16.143 từ và 17.756 synsets. Phương pháp này hoàn toàn có thể được áp dụng để xây dựng WordNet cho các ngôn ngữ khác, cũng như thêm vào WordNet của các ngôn ngữ nguồn để tăng số lượng từ và synsets được xây dựng [2]. Năm 2022, Lâm và cộng sự đã xây dựng Mạng từ tiếng Việt (VWN) từ các nguồn miễn phí của nhiều ngôn ngữ khác nhau. VWN được xây dựng có cấu trúc tương tự như Mạng từ tiếng Anh (Princeton WordNet). Nhóm tác

giả đã sử dụng thuật toán tự động để chuyển ngữ các từ tiếng Anh trong WordNet sang tiếng Việt. Nhóm tác giả đã loại bỏ sự nhập nhằng của kết quả dịch bằng cách xếp hạng dựa trên số lần xuất hiện và khoảng cách của Google đối với các từ trong bản dịch, đồng thời kết nối giữa các tập đồng nghĩa (synsets) và trích xuất các định nghĩa (glosses) cho synsets. Nhóm tác giả cũng đã xem xét các vấn đề gặp phải trong VWN do sự khác biệt về văn hóa và hình thái đơn lập của tiếng Việt so với các ngôn ngữ khác [7]. Năm 2023, nghiên cứu của Trang T. M. P. và cộng sự về tính toán sự tương đồng ngữ nghĩa giữa hai danh từ tiếng Việt dựa trên mạng từ tiếng Việt. Nghiên cứu tập trung vào việc đo khoảng cách ngữ nghĩa giữa các danh từ tiếng Việt để xác định sự tương đồng và dị biệt của chúng từ góc độ ngữ nghĩa. Nghiên cứu sử dụng một tập ngữ liệu gồm 1000 cặp danh từ được lựa chọn ngẫu nhiên từ mạng từ tiếng Việt và áp dụng các độ đo khoảng cách ngữ nghĩa khác nhau như Path-length, Leacock & Chodorow, Wu & Palmer, Resnik, Lin, và Jiang - Conrath để đo khoảng cách ngữ nghĩa của các cặp danh từ trên 3 cấp độ (đo cùng lớp ý niệm nguyên thủy, đo 2 lớp ý niệm nguyên thủy gần nhau và đo 2 lớp ý niệm nguyên thủy xa nhau). Kết quả cho thấy độ đo khoảng cách ngữ nghĩa Resnik phù hợp trong đánh giá sự tương đồng giữa các ý niệm [12].

3. Tổ chức mạng từ

3.1. Tổ chức mạng từ WordNet

Phiên bản trực tuyến mới nhất hiện nay của WordNet là phiên bản 3.1, với hơn 155.000 từ. Trong WordNet, hệ thống ý niệm tiếng Anh được tổ chức thành các nhóm từ có quan hệ ngữ nghĩa với nhau, được biểu diễn dưới dạng các tập đồng nghĩa (synsets). Các danh từ trong WordNet được tổ chức thành 11 nhóm ý niệm nguyên thủy (primitive concepts), trong đó lớn nhất là *ý niệm thực thể* (entity). Trong ý niệm thực thể (entity), WordNet chia thành hai loại chính: *vật có sự sống* (living thing) và *vật thể không có sự sống* (non-living thing). Trong vật có sự sống, có ba danh mục chính: *thực vật* (plant), *động vật* (animal) và *con người* (person). Trong vật thể không có sự sống, có bốn danh mục chính: *vật thể tự nhiên* (natural object), *đồ nhân tạo* (artifact), *chất* (substance) và *thức ăn* (food). Hình 1 tổ chức và phân loại các ý niệm nguyên thủy của các danh từ thực thể trong một cấu trúc hệ thống, giúp hiểu rõ hơn về cách phân loại và tư duy về các loại đối tượng khác nhau trong thế giới xung quanh chúng ta.



Hình 1. Sơ đồ ý niệm nguyên thủy của danh từ thực thể (entity) [8]

3.2. Tổ chức mạng từ VietNet

Với sự phát triển của ngôn ngữ học tính toán và nhu cầu trong việc xử lý ngôn ngữ tiếng Việt, một hệ thống ý niệm Tiếng Việt đã được xây dựng bởi Trung tâm Ngôn ngữ học Tính toán (CLC) có tên gọi là VietNet. VietNet được xây dựng dựa trên sự so sánh tương đồng giữa định nghĩa trong cấu trúc mạng từ tiếng Anh (Princeton WordNet) và định nghĩa trong từ điển OALD (Oxford Advanced Learner's Dictionary) with Vietnamese Translation, ấn bản lần thứ 8 [6]. Nếu hai định nghĩa đó tương đồng nhau thì sẽ lấy các từ tiếng Việt tương đương được dịch ra từ các từ tiếng Anh trên.

VietNet có 21.190 danh từ với 47.300 mục từ. Trong đó danh từ cũng được chia thành 11 nhóm ý niệm nguyên thủy tương tự như WordNet. Tuy nhiên, thành phần mỗi nhánh có thể khác nhau. Do đó, để đánh giá sự tương đồng về mặt cấu trúc, chúng ta dựa vào độ đo khoảng cách ngữ nghĩa.

3.3. Độ đo Resnik

Khoảng cách ngữ nghĩa là một khái niệm trong ngôn ngữ học tính toán, đo lường mức độ gần hoặc xa về mặt ngữ nghĩa giữa các đơn vị ngôn ngữ như từ, cụm từ, câu, đoạn văn, tài liệu... dựa trên ý nghĩa của chúng (Mohammad, 2008) [9]. Đo khoảng cách ngữ nghĩa giúp chúng ta định lượng được mức độ tương đồng hoặc dị biệt ngữ nghĩa giữa các đơn vị ngôn ngữ.

Có nhiều độ đo khoảng cách ngữ nghĩa, trong đó độ đo Resnik được sử dụng phổ biến trong việc tính toán khoảng cách ngữ nghĩa giữa các đơn vị ngôn ngữ. Độ đo Resnik, được xây dựng năm 1995, thuộc nhóm giải thuật sử dụng nội dung thông tin. Độ đo Resnik được tính bằng cách sử dụng thông tin về định nghĩa tập đồng nghĩa (synset) của hai từ. Độ đo này đo tương đồng từ không chỉ sử dụng cấu trúc của mạng WordNet mà còn dựa vào nội dung thông tin, xác suất thông tin xuất hiện trong kho ngữ liệu [11].

Nội dung thông tin (IC: Information Content) của ý niệm

Công thức nội dung thông tin:

$$IC(c) = -\log P(c) \quad (1)$$

Công thức tính xác suất của ý niệm c :

$$P(c) = \frac{\sum_{w \in words(c)} count(w)}{N} \quad (2)$$

- Trong đó:
- w là những từ cùng mang ý niệm c .
 - N là tổng số từ có trong kho ngữ liệu.

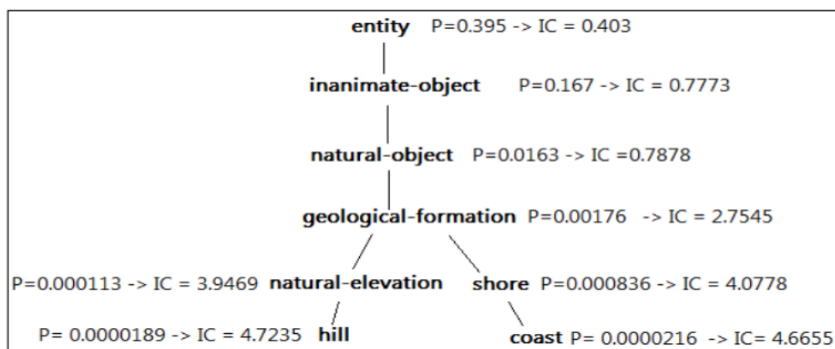
Độ đo tương đồng Resnik của hai ý niệm được tính bằng lượng nội dung thông tin của điểm nút cha chung gần nhất:

$$simResnik(c1, c2) = IC(LCS(c1, c2)) \quad (3)$$

Ví dụ theo cây phân cấp như Hình 2 thì độ đo Resnik giữa hai từ “hill” và “shore” là:

$$simResnik(hill, shore) = IC(geological - formation) = 2,7545$$

Bên cạnh đó, độ đo Resnik có đặc điểm là *simResnik* (hill, shore) = *sim* (hill, coast) vì chúng có cùng cha chung.



Hình 2. Nội dung thông tin trên cây WordNet

4. Phương pháp nghiên cứu

Để thực hiện nghiên cứu, chúng tôi tiến hành thu thập ngữ liệu về mạng ý niệm danh từ tiếng Việt và tiếng Anh.

4.1. Thu thập ngữ liệu

Đầu tiên, chúng tôi thu thập ngữ liệu là các danh từ thực thể từ hai nguồn chính: tiếng Việt từ nguồn VietNet và tiếng Anh từ nguồn WordNet. Để tiến hành phân tích và so sánh cấu trúc của VietNet và WordNet trên các danh từ thực thể, chúng tôi đã thu thập một tập ngữ liệu đa dạng và đại diện gồm 1.400 cặp danh từ thực thể tiếng Việt và 1.400 cặp danh từ thực thể tiếng Anh. Ngữ liệu đã được lựa chọn dựa trên bảy ý niệm con của ý niệm danh từ thực thể, mỗi ý niệm con gồm 200 cặp danh từ. Các ý niệm con bao gồm: “Thực vật” (Plant), “Động vật” (Animal), “Con người” (Person), “Vật thể tự nhiên” (Natural object), “Đồ nhân tạo” (Artifact), “Chất” (Substance) và “Thức ăn” (Food).

4.2. Tính toán khoảng cách ngữ nghĩa

Chúng tôi áp dụng độ đo Resnik cho từng cặp danh từ thực thể tiếng Việt và tiếng Anh trong ngữ liệu thu thập. Kết quả của độ đo Resnik là một giá trị khoảng cách ngữ nghĩa, đo lường mức độ tương đồng giữa hai từ. Sau khi áp dụng độ đo Resnik cho các cặp danh từ thực thể tiếng Việt và tiếng Anh, chúng tôi tính toán khoảng cách ngữ nghĩa trên bảy ý niệm con của ý niệm nguyên thủy danh từ thực thể và thu được giá trị tương ứng cho mỗi cặp. Chúng tôi sử dụng phương pháp thống kê mô tả (descriptive statistics) để thu thập và phân tích số liệu từ các kết quả đo khoảng cách ngữ nghĩa. Bằng cách sử dụng các thống kê như giá trị nhỏ nhất (min), giá trị lớn nhất (max), trung bình (mean), độ lệch chuẩn (Std. Deviation) ..., từ đó có thể mô tả các đặc điểm ngôn ngữ liên quan đến khoảng cách ngữ nghĩa. Bên cạnh đó, phương pháp định lượng (quantitative method) được sử dụng để đo lường và số hóa các kết quả đo khoảng cách ngữ nghĩa. Bằng cách áp dụng các phép đo và chỉ số số học, chúng tôi có thể tạo ra các giá trị định lượng để so sánh và đánh giá kết quả.

4.3. So sánh cấu trúc

Chúng tôi so sánh giá trị khoảng cách ngữ nghĩa giữa các cặp danh từ thực thể tiếng Việt và tiếng Anh. Chênh lệch giữa các giá trị khoảng cách ngữ nghĩa cho ta biết mức độ tương đồng cấu trúc giữa VietNet và WordNet. Phương pháp so sánh - đối chiếu (comparative method) được sử dụng để so sánh kết quả đo khoảng cách ngữ nghĩa danh từ tiếng Việt với tiếng Anh. Bằng cách so sánh và phân tích sự tương quan và khác biệt giữa các kết quả, đánh giá cấu trúc của VietNet và WordNet từ góc độ ngữ nghĩa, mở ra những cơ hội và tiềm năng trong việc nghiên cứu và ứng dụng ngôn ngữ học máy tính cho tiếng Việt.

4.4. Phân tích kết quả

Chúng tôi trình bày và phân tích kết quả để đánh giá mức độ tương đồng cấu trúc giữa VietNet và WordNet trên các danh từ thực thể. Bên cạnh đó, chúng tôi cũng đánh giá hiệu quả của việc áp dụng độ đo Resnik cho các cặp danh từ thực thể tiếng Việt và tiếng Anh. Những phát hiện này có ý nghĩa quan trọng và ứng dụng trong nhiều lĩnh vực của ngôn ngữ học, như đối chiếu Anh - Việt về mặt ngữ nghĩa học từ vựng hay tương đồng văn bản xuyên ngôn ngữ.

5. Kết quả nghiên cứu và thảo luận

Từ kết quả nghiên cứu trên, chúng tôi nhận thấy có sự tương đồng đáng kể trong cấu trúc giữa VietNet và WordNet trên các danh từ thực thể. Bằng cách sử dụng độ đo tương đồng ngữ nghĩa Resnik, chúng tôi đã tính toán khoảng cách ngữ nghĩa cho 1.400 cặp danh từ thực thể tiếng Việt và 1400 cặp danh từ thực thể tiếng Anh. Bảng 1, Bảng 3 và Bảng 5 cung cấp các thông tin về giá trị khoảng cách ngữ nghĩa trung bình của danh từ thực thể tiếng Việt (VietNet), danh từ thực thể tiếng Anh (WordNet) và cung cấp các chỉ số liên quan để so sánh phân tích kết quả.

Kết quả bảng 1 cung cấp cho chúng ta một cái nhìn tổng quan về khoảng cách ngữ nghĩa trung bình của các danh từ thực thể trong VietNet, cho thấy sự tương đồng và dị biệt giữa các danh từ thực thể, đồng thời chỉ ra mức độ biến động và phân tán của khoảng cách ngữ nghĩa.

Bảng 1. Giá trị khoảng cách ngữ nghĩa trung bình của danh từ thực thể tiếng Việt (VietNet)

Danh từ thực thể (VietNet)	Thực vật	Động vật	Con người	Vật thể tự nhiên	Đồ nhân tạo	Chất	Thức ăn	Trung bình tổng
Trung bình (Mean)	3,7906	3,9079	2,5291	3,8187	2,4863	3,6915	4,9208	3,5921
Giá trị nhỏ nhất (Min)	-,0336	-0,0336	-,0336	-0,0336	-0,0336	0,5181	-,0336	0,0452
Giá trị lớn nhất (Max)	9,5571	11,1766	8,4118	8,0807	7,5620	8,6591	10,7961	9,1776
Phương sai (Variance)	1,6850	3,0847	1,0988	2,5660	1,1261	2,3849	4,2554	2,3144
Độ lệch chuẩn (Std. Deviation)	1,2981	1,7563	1,0482	1,6019	1,0612	1,5443	2,0629	1,4818

Trung bình của khoảng cách ngữ nghĩa cho mỗi loại danh từ thực thể trong tiếng Việt cho thấy rằng danh từ chỉ “đồ nhân tạo” có giá trị trung bình thấp nhất, trong khi đó danh từ chỉ “thức ăn” có giá trị trung bình cao nhất. Điều này cho thấy rằng danh từ chỉ “đồ nhân tạo” có sự tương đồng hơn với các danh từ thực thể khác trong VietNet, trong khi đó danh từ chỉ “thức ăn” có sự khác biệt nhiều hơn.

Các giá trị phương sai và độ lệch chuẩn cho thấy mức độ biến động và phân tán của khoảng cách ngữ nghĩa giữa các danh từ thực thể. Sự khác biệt trong phương sai và độ lệch chuẩn giữa các loại trong danh từ thực thể cho thấy rằng có sự đa dạng trong mức độ tương đồng của chúng. Các danh từ chỉ “thức ăn” và danh từ chỉ “động vật” có phương sai và độ lệch chuẩn cao, tức là có sự biến động và phân tán lớn trong khoảng cách ngữ nghĩa của chúng.

Bảng 2 trình bày giá trị khoảng cách ngữ nghĩa trung bình của một số cặp danh từ điển hình thuộc ý niệm “thực thể” trong tiếng Việt (VietNet).

Bảng 2. Giá trị khoảng cách ngữ nghĩa trung bình của một số cặp danh từ thực thể trong tiếng Việt (VietNet)

Danh từ thực thể (VietNet)	Một số cặp danh từ điển hình	Giá trị khoảng cách ngữ nghĩa trung bình
Thực vật	dâm bụt – cây tía tô đất	6,3482
	tỏi – rau càng cua	3,7991
Động vật	sâu róm – quạ chân đỏ	4,0108
	cáo túi – linh dương	5,2369
Con người	ngư dân – người bán vé	5,4638
	nhà soạn kịch – công nhân chuyển hàng	2,4662
Vật thể tự nhiên	quả nang – củ từ	3,8198
	quả khế – quả mận hoang	6,7002
Đồ nhân tạo	đồ hóa trang – đèn đọc sách	2, 2190
	gương trang điểm – cửa	2,2190
Chất	sáp thực vật – nhựa cây	2,5229
	sợi tự nhiên – chất tổng hợp	3,3346
Thức ăn	cháo ngô – cao lương	2,2818
	thịt nhồi – súp gà	4,8610

Với ý niệm “thực vật”, khoảng cách giữa “dâm bụt” và “cây tía tô đất” là 6,3482, còn giữa “tỏi” và “rau càng cua” là 3,7991. Về ý niệm “động vật”, khoảng cách giữa “sâu róm” và “quạ chân đỏ” là 4,0108, giữa “cáo túi” và “linh dương” là 5,2369. Đối với ý niệm “con người”, khoảng cách giữa “ngư dân” và “người bán vé” là 5,4638, còn giữa “nhà soạn kịch” và “công nhân chuyển hàng” là 2,4662. Với vật thể tự nhiên, khoảng cách giữa “quả nang” và “củ từ”

là 3,8198, giữa “quả khế” và “quả mận hoang” là 6,7002. Với ý niệm “đồ nhân tạo”, khoảng cách giữa “đồ hóa trang” và “đèn đọc sách” là 2,2190, giữa “gương trang điểm” và “cửa” cũng là 2,2190. Về chất, khoảng cách giữa “sáp thực vật” và “nhựa cây” là 2,5229, còn giữa “sợi tự nhiên” và “chất tổng hợp” là 3,3346. Với ý niệm thức ăn, khoảng cách giữa “cháo ngô” và “cao lương” là 2,2818, còn giữa “thịt nhồi” và “súp gà” là 4,8610.

Kết quả Bảng 3 cung cấp thông tin về khoảng cách ngữ nghĩa trung bình của các danh từ thực thể trong WordNet. Kết quả phân tích Bảng 3 cung cấp cái nhìn tổng quan về mức độ tương đồng và dị biệt của các danh từ thực thể trong WordNet dựa trên khoảng cách ngữ nghĩa trung bình. Các giá trị thống kê như trung bình, giá trị nhỏ nhất, giá trị lớn nhất, phương sai và độ lệch chuẩn giúp chúng ta hiểu rõ hơn về sự đa dạng và biến động trong ngữ nghĩa của các danh từ thực thể tiếng Anh.

Bảng 3. Giá trị khoảng cách ngữ nghĩa trung bình của danh từ thực thể tiếng Anh (WordNet)

Danh từ thực thể (WordNet)	Plant (Thực vật)	Animal (Động vật)	Person (Con người)	Natural object (Vật thể tự nhiên)	Artifact (Đồ nhân tạo)	Substance (Chất)	Food (Thức ăn)	Trung bình tổng
Trung bình (Mean)	4,1680	4,2518	2,2577	3,6411	2,3199	2,9009	3,7976	3,3338
Giá trị nhỏ nhất (Min)	-0,0552	-0,0552	-0,0552	-0,0552	-0,0552	-0,0552	-0,0552	-0,0552
Giá trị lớn nhất (Max)	9,1061	10,6166	8,2084	13,6612	12,1645	9,0493	9,6556	10,3517
Phương sai (Variance)	2,0534	3,1823	1,1854	4,4092	2,0301	2,4235	5,4756	2,9656
Độ lệch chuẩn (Std. Deviation)	1,4330	1,7839	1,0887	2,0998	1,4248	1,5568	2,3400	1,6753

Thông qua kết quả từ Bảng 3, chúng ta có thể nhận thấy rằng giá trị trung bình cho danh từ chỉ “con người” là thấp nhất, trong khi danh từ chỉ “động vật” có giá trị trung bình cao nhất. Điều này cho thấy rằng danh từ chỉ “con người” có sự tương đồng hơn với các danh từ thực thể khác trong WordNet, trong khi đó danh từ chỉ “động vật” có sự khác biệt nhiều hơn.

Ngoài ra, chúng ta cũng có thể quan sát được mức độ biến động và phân tán của khoảng cách ngữ nghĩa. Các giá trị phương sai và độ lệch chuẩn cho thấy mức độ đa dạng và sự chênh lệch trong khoảng cách ngữ nghĩa của các danh từ. Các danh từ chỉ “thức ăn” và danh từ chỉ “vật thể tự nhiên” có phương sai và độ lệch chuẩn cao, tức là có sự biến động và phân tán lớn trong khoảng cách ngữ nghĩa của chúng.

Bảng 4 trình bày giá trị khoảng cách ngữ nghĩa trung bình của một số cặp danh từ điển hình thuộc ý niệm “thực thể” trong tiếng Anh (WordNet).

Bảng 4. Giá trị khoảng cách ngữ nghĩa trung bình của một số cặp danh từ thực thể trong tiếng Anh (WordNet)

Danh từ thực thể (WordNet)	Một số cặp danh từ điển hình	Giá trị khoảng cách ngữ nghĩa trung bình
Plant (Thực vật)	aloe (nha đam) – apple (táo)	1,5162
	ancient pine (cây thông cổ) – hollyhock (cây thực quỳ)	5,1085
Animal (Động vật)	marlin (cá tuyết than) – sunfish (cá thái dương)	7,4862
	driftfish (cá chim) – bison (bò rừng bison)	4,5544
Person (Con người)	footman (người hầu) – procurator (kiểm sát viên)	4,9059
	sniper (kẻ bắn tỉa) – waker (người đánh thức)	2,4220
Natural object (Vật thể tự nhiên)	nutlet (quả hạch) – lodestar (sao bắc cực)	-0,0552
	alula (vây gốc cánh) – snowdrift (đống tuyết)	4,3830
Artifact (Đồ nhân tạo)	strickle (que gạt) – plate (đĩa)	2,4848
	answering machine (máy trả lời) – circuit (mạch điện)	4,3741
Substance (Chất)	benzoate (chất bảo quản thực phẩm benzoate) – margarin (bơ thực vật)	3,3005
	carnelian (khoáng chất carnelian – seawater (nước biển)	3,0731
Food (Thức ăn)	cranberry sauce (nước sốt việt quất) – yolk (lòng đỏ)	3,5058
	canape (bánh canape) – curry (cà ri)	4,0016

Với ý niệm “thực vật”, khoảng cách giữa “aloe” (nha đam) và “apple” (táo) là 1,5162, còn giữa “ancient pine” (cây thông cổ) và “hollyhock” (cây thực quỳ) là 5,1085. Về ý niệm “động vật”, khoảng cách giữa “marlin” (cá tuyết than) và “sunfish” (cá thái dương) là 7,4862, còn giữa “driftfish” (cá chim) và “bison” (bò rừng bison) là 4,5544. Đối với ý niệm “con người”, khoảng cách giữa “footman” (người hầu) và “procurator” (kiểm sát viên) là 4,9059, trong khi khoảng cách giữa “sniper” (kẻ bắn tỉa) và “waker” (người đánh thức) là 2,4220. Với ý niệm “vật thể tự nhiên”, khoảng cách giữa “nutlet” (quả hạch) và “lodestar” (sao bắc cực) là -0,0552, còn giữa “alula” (vây gốc cánh) và “snowdrift” (đống tuyết) là 4,3830. Trong ý niệm “đồ nhân tạo”, khoảng cách giữa “strickle” (que gạt) và “plate” (đĩa) là 2,4848, còn giữa “answering machine” (máy trả lời) và “circuit” (mạch điện) là 4,3741. Đối với ý niệm “chất”, khoảng cách giữa “benzoate” (chất bảo quản thực phẩm benzoate) và “margarin” (bơ thực vật) là 3,3005, còn giữa “carnelian” (khoáng chất carnelian) và “seawater” (nước biển) là 3,0731. Với ý niệm “thức ăn”, khoảng cách giữa “cranberry sauce” (nước sốt việt quất) và “yolk” (lòng đỏ) là 3,5058, còn giữa “canape” (bánh canape) và “curry” (cà ri) là 4,0016.

Bảng 5 so sánh giá trị khoảng cách ngữ nghĩa trung bình của danh từ thực thể trong tiếng Việt (VietNet) và danh từ thực thể trong tiếng Anh (WordNet) cũng như so sánh sự chênh lệch trung bình khoảng cách ngữ nghĩa giữa các danh từ thực thể trong hai ngôn ngữ này.

Bảng 5. So sánh giá trị khoảng cách ngữ nghĩa trung bình của danh từ thực thể tiếng Việt (VietNet) và tiếng Anh (WordNet)

Danh từ thực thể (Entity)	Khoảng cách ngữ nghĩa danh từ tiếng Việt (VietNet)	Khoảng cách ngữ nghĩa danh từ tiếng Anh (WordNet)	Chênh lệch trung bình khoảng cách ngữ nghĩa danh từ tiếng Việt và tiếng Anh
Thực vật (Plant)	3,7906	4,1680	0,3774
Động vật (Animal)	3,9079	4,2518	0,3439
Con người (Person)	2,5291	2,2577	0,2714
Vật thể tự nhiên (Natural object)	3,8187	3,6411	0,1776
Đồ nhân tạo (Artifact)	2,4863	2,3199	0,1664
Chất (Substance)	3,6915	2,9009	0,7906
Thức ăn (Food)	4,9208	3,7976	1,1232
Trung bình khoảng cách	3,5921	3,3338	0,2583

Theo nguyên lí tổ chức phân cấp ý niệm chung của các mạng từ như WordNet, VietNet..., mức độ tương đồng giữa các danh từ trong hai ngôn ngữ sẽ càng cao khi giá trị chênh lệch khoảng cách trung bình giữa hai ngôn ngữ càng nhỏ (về 0). Nhìn chung, kết quả so sánh cho thấy giá trị khoảng cách ngữ nghĩa giữa các danh từ thực thể tiếng Việt và tiếng Anh tương đối gần nhau, với chênh lệch nhỏ là 0,2583. Điều này cho thấy cấu trúc của VietNet đáng tin cậy đối với các danh từ thực thể, vì nó có sự tương đồng cao với cấu trúc của WordNet. Khi xem xét từng loại danh từ cụ thể, ta có thể nhận thấy có sự khác biệt trong ngữ nghĩa của các danh từ thực thể giữa tiếng Việt và tiếng Anh. Ví dụ, danh từ chỉ “thức ăn” (food) có chênh lệch trung bình khoảng cách ngữ nghĩa là 1,1232 cho thấy có tồn tại sự khác biệt trong ngữ nghĩa của danh từ này giữa hai ngôn ngữ. Tương tự, danh từ chỉ “chất” (substance) có chênh lệch trung bình khoảng cách ngữ nghĩa là 0,7906 cũng chỉ ra có tồn tại sự khác biệt trong ngữ nghĩa của loại danh từ này giữa hai ngôn ngữ. Tuy nhiên, chênh lệch trung bình của các danh từ khác như “thực vật” (plant), “động vật” (animal), “con người” (person), “vật thể tự nhiên” (natural object) và “đồ nhân tạo” (artifact) có chênh lệch trung bình khoảng cách ngữ nghĩa nhỏ hơn chỉ từ 0,1664 đến 0,3774. Điều này cho thấy có sự tương đồng đáng kể trong ngữ nghĩa của các loại danh từ này trong danh từ thực thể giữa hai ngôn ngữ.

Xem xét chi tiết hơn các giá trị khoảng cách ngữ nghĩa trung bình cho mỗi loại danh từ của danh từ thực thể trong tiếng Việt và tiếng Anh ta thấy một số giá trị trung bình khoảng cách ngữ nghĩa của các danh từ trong VietNet có xu hướng cao hơn so với WordNet và ngược lại cũng có một số giá trị trung bình khoảng cách ngữ nghĩa của các danh từ trong WordNet cao hơn so với VietNet. Ví dụ, giá trị trung bình khoảng cách ngữ nghĩa của các danh từ chỉ “thức ăn” trong VietNet (4,9208) có xu hướng cao hơn so với WordNet (3,7976). Lấy một ví dụ khác về danh từ chỉ “thực vật” (plant), giá trị trung bình khoảng cách ngữ nghĩa của các danh từ “thực vật” trong WordNet (4,1680) có xu hướng cao hơn so với VietNet (3,7906).

6. Kết luận

Qua kết quả nghiên cứu về ngữ nghĩa của các danh từ thực thể trong tiếng Việt (VietNet) và tiếng Anh (WordNet) giúp ta hiểu được sự tương đồng và dị biệt giữa hai ngôn ngữ này. Kết quả nghiên cứu làm nổi bật tiềm năng của VietNet và độ đo Resnik trong việc so sánh giá

trị khoảng cách ngữ nghĩa của các danh từ giữa tiếng Việt và tiếng Anh. Bên cạnh đó, kết quả này cũng góp phần vào việc mở rộng kiến thức về độ đo tương đồng ngữ nghĩa giữa các ngôn ngữ và làm cơ sở cho những nghiên cứu tiếp theo.

Các phát hiện từ nghiên cứu này có ý nghĩa quan trọng và đem lại nhiều ứng dụng tiềm năng trong lĩnh vực của ngữ nghĩa học từ vựng và ngôn ngữ học tính toán như: xây dựng tự động từ điển Anh-Việt (giúp thu thập ngữ cảnh sử dụng từ), phát triển các công cụ hỗ trợ dịch thuật thông minh (giúp nâng cao hiệu quả và chất lượng của quá trình dịch thuật), đánh giá tương đồng văn bản xuyên ngôn ngữ (hỗ trợ việc so sánh và phân tích nội dung giữa các văn bản tiếng Việt và tiếng Anh), nghiên cứu ngôn ngữ và văn hóa (giúp hiểu rõ hơn về cấu trúc và tương đồng ngữ nghĩa giữa các ngôn ngữ khác nhau)...

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Đinh Điền, *Sách chuyên khảo ngôn ngữ học ngữ liệu*, Nxb Đại học Quốc gia TP. HCM, 2018.
2. Trương Hoàng Bảo, Nguyễn Ngọc Bình, Lê Thành Nguyên và cộng sự, *Xây dựng WordNet tiếng Việt tự động bằng ngữ liệu song ngữ*, in Kỷ yếu Hội nghị KH-CN Quốc gia lần thứ XIV về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Trường Đại học Khoa học Tự nhiên, ĐHQG TP. HCM, ngày 23-24/12/2021, 483-490, 2021.

Tiếng Anh

3. Duc, H. N., & Thao, N. T., *Towards building a WordNet for Vietnamese*, In Proceedings of the 1st International Workshop for Computer, Information and Communication Technologies, Hanoi, Vietnam, 2003.
4. Duong, T. H., Tran, M. Q., & Nguyen, T. P. T., *Collaborative Vietnamese WordNet building using consensus quality*, Vietnam Journal of Computer Science, 4, 85-96, 2017.
5. Guo, X., Zhu, X., Li, F., & Li, Q., *A new semantic similarity measurement based on HowNet concept tree*, In 2016 International Forum on Management, Education and Information Technology Application (pp. 722-727), Atlantis Press, January, 2016.
6. Joanna, T. & Dien, D., *Oxford Advanced Learner's Dictionary* 8th Edition (With Vietnamese Translation), Tre Publishing House, 2014.
7. Lam, K. N., & Kalita, J., *Constructing Vietnamese Wordnet: A case study*, Computación y Sistemas, 26(3), 1313-1322, 2022.
8. Miller, G.A., et al., *Introduction to WordNet: An on-line lexical database*, International journal of lexicography, 3(4): p. 235-244, 1990.
9. Mohammad, S., *Measuring semantic distance using distributional profiles of concepts*. University of Toronto, 2008.
10. Nguyen, T. P., Pham, V. L., Nguyen, H. A., Vu, H. H., & Tran, N. A., *A two-phase approach for building Vietnamese wordnet*, In Proceedings of the 8th Global WordNet Conference (GWC) (pp. 261-266), 2016.
11. Resnik, P., *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*, In Proceeding, IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1, pages: 448-453, 1995.
12. Trang, T. M. P., An, T. D., & Phuong, T. M. T., *WordNet-based Computation of Semantic Similarity Between Two Vietnamese Nouns*, Technium Social Sciences journal, 43, 532-545, 2023.
13. Vossen, P., *WordNet, EuroWordNet and Global WordNet*, Revue française de linguistique appliquée, 7(1), 27-38, 2002.