ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Bùi Văn Tân

TỰ ĐỘNG XÁC ĐỊNH QUAN HỆ NGỮ NGHĨA CỦA TỪ DỰA TRÊN HỌC MÁY THỐNG KÊ

TÓM TẮT LUẬN ÁN TIẾN SỸ CÔNG NGHỆ THÔNG TIN

Hà Nội - 2021

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Bùi Văn Tân

TỰ ĐỘNG XÁC ĐỊNH QUAN HỆ NGỮ NGHĨA CỦA TỪ DỰA TRÊN HỌC MÁY THỐNG KÊ

Chuyên ngành: Khoa học máy tính
Mã số: 9480101.01

TÓM TẮT LUẬN ÁN TIẾN SỸ CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. Nguyễn Phương Thái

Hà Nội - 2021

Mục lục

Chương 1

GIỚI THIỆU

1.1 Đặt vấn đề

Lĩnh vực nghiên cứu xử lý ngôn ngữ tự nhiên (NLP) nhắm đến mục đích làm cho máy (như: *máy tính, robot, thiết bị thông minh,...*) có thể hiểu được ngôn ngữ tự nhiên của con người và hiện nay, khả năng này của máy vẫn còn ở mức độ hạn chế. Những năm gần đây, lĩnh vực nghiên cứu về NLP xuất hiện những hướng tiếp cận mới "mạnh mẽ" như học máy, học sâu. Một trong những chủ đề nghiên cứu quan trọng trong NLP là *xác định quan hệ ngữ nghĩa của từ vựng* (Lexical Semantic Relation Determination). Trong những năm gần đây bài toán này đã nhận được sự quan tâm nghiên cứu đặc biệt cũng như có nhiều kết quả nghiên cứu được công bố.

Trong luận án này, chúng tôi sử dụng tiếp cận ngữ nghĩa phân phối dựa trên mạng nơron, khai thác một số mô hình học máy, học sâu để xác định một số quan hệ ngữ nghĩa của cặp từ gồm quan hệ bao thuộc (Hypernymy), đồng nghĩa (Synonymy), trái nghĩa (Antonymy), tương đồng ngữ nghĩa (Semantic Similarity). Cụ thể, luận án này nhắm đến giải quyết ba bài toán gồm: Xác định quan hệ bao thuộc (Hypernymy Recognition - HR); phân tách các cặp từ theo quan hệ đồng nghĩa-trái nghĩa (Antonymy-Synonymy Classification - ASC); đo lường độ tương tự ngữ nghĩa của cặp từ (Word Similarity Measurement - WSM). Bảng 1.1 trình bày tóm tắt về đặc trưng của các bài toán cũng như các quan hệ ngữ nghĩa, ngôn ngữ được nghiên cứu trong luận án này.

Bảng 1.1: Đặc trưng của các bài toán về loại kết quả đầu ra, quan hệ ngữ nghĩa, và ngôn ngữ, được nghiên cứu trong luận án.

Bài toán	Đặc trưng đầu ra		Đặc trưng quan hệ		Đặc trưng ngôn ngữ	
	Định tính	Định lượng	Đối xứng	Bất đối xứng	Tiếng Anh	Tiếng Việt
HR	✓			✓	✓	✓
ASC	✓		✓	✓		✓
WSM		✓	✓	✓	✓	✓

1.2 Hướng tiếp cận và phương pháp

Đối với bài toán phát hiện quan hệ bao thuộc, chúng tôi sử dụng hai chiến lược chính. Thứ nhất, chúng tôi nhắm đến mục tiêu học được các biểu diễn vector "chuyên biệt" cho từ (Specialized Word Embeddings) bằng một mô hình mạng nơron. Biểu diễn vector cho chuyên biệt này không những chứa thông tin ngữ nghĩa của từ như các biểu diễn nhúng từ thông thường (Word2vec, GloVe, fastText...) mà còn mã hóa những đặc trưng của quan hệ bao thuộc. Thứ hai, chúng tôi nhắm đến khai thác các đặc trưng về cấu trúc của các từ ghép và thuật ngữ. Bằng cách kết hợp vector nhúng từ chuyên biệt với vector đặc trưng cấu trúc của từ, mô hình được đề xuất trong luận án đã cải thiện hiệu năng đáng kể cho bài toán so với các mô hình đã đề xuất trước đó.

Đối với bài toán phân tách các cặp từ có quan hệ đồng nghĩa, trái nghĩa, chúng tôi sử dụng hai chiến lược chính. Thứ nhất, chúng tôi khai thác thông tin ngữ cảnh đồng hiện của cặp từ. Thông tin đồng hiện của cặp từ được mã hóa thành vector bởi một mô hình mạng nơron. Thứ hai, chúng tôi dử dụng một số đặc trưng quan trọng giúp phân tách quan hệ đồng nghĩa, trái nghĩa như thông tin tương hỗ theo từng cặp, đặc trưng về mối quan hệ ngữ nghĩa giữa các thành phần của từ này với các thành phần của từ kia trong một cặp từ. Đối với bài toán phát hiện quan hệ trái nghĩa, chúng tôi sử dụng các cặp từ có quan hệ đồng nghĩa, trái nghĩa trích được từ WordNet và từ điển để học các biểu diễn vector chuyên biệt cho từ. Những vector nhúng từ chuyên biệt đã được mã hóa thêm các thông tin về quan hệ đồng nghĩa, trái nghĩa. Thêm nữa, thông tin về độ đo thông tin tương hỗ của cặp từ cũng được khai thác để tăng hiệu năng của mô hình.

Đối với bài toán đo lường độ tương tự của cặp từ, chúng tôi đề xuất những cải tiến nhằm tăng độ chính xác trong đo lường độ tương tự của cặp từ đơn ngữ và song ngữ. Để lượng giá chính xác hơn độ tương tự của cặp từ, chiến lược thứ nhất là áp dụng thuật toán tìm đường đi tối ưu giữa hai đỉnh của đồ thị để đo khoảng cách ngữ nghĩa ngắn nhất giữa hai từ. Chiến lược thứ hai, chúng tôi khai thác thông tin định nghĩa của các từ. Chúng tôi giải thiết rằng, độ tương tự ngữ nghĩa giữa hai từ tương quan với mức độ tương đồng ngưa nghĩa giữa các định nghĩa của chúng.

1.3 Các kết quả chính của luận án

Trong luận án này, chúng tôi hướng đến nâng cao hiệu năng của các mô hình tự động xác định một số quan hệ ngữ nghĩa của từ theo tiếp cận học máy thống kê gồm: quan hệ bao thuộc, quan hệ đồng nghĩa-trái nghĩa, quan hệ tương đồng ngữ nghĩa.

Đối với bài toán phát hiện quan hệ bao thuộc (Hypernymy Recognition - HR), luận án đã đề xuất một mô hình mạng nơron học các vector nhúng từ chuyên biệt (specialized word embedding vector). Các vector nhúng từ học được phù hợp cho bài toán phát hiện quan hệ hypernymy hơn các mô hình nhúng từ đã được đề xuất trước đó như Word2Vec, fastText, GloVe. Những đặc trưng của từ ghép có thể là những dấu hiệu quan trọng

giúp nhận ra quan hệ hyernymy của cặp từ. Luận án đã đề xuất một lược đồ trích chọn những đặc trưng ngữ nghĩa mức dưới từ (Subword Semantic Feature). Lược đồ được đề xuất không những mã được quan hệ ngữ nghĩa của các thành phần của cặp từ mà còn nắm bắt được cả thông tin vị trí của chúng trong các vector đặc trưng ngữ nghĩa dưới từ. Để phát hiện các cặp từ có quan hệ hypernymy, mô hình phân lớp có giám sát Support Vector Machine được sử dụng với đặc trưng đầu vào được kết hợp từ vector nhúng từ và vector đặc trưng ngữ nghĩa dưới từ. Kết quả thực nghiệm được đánh giá trên một số bộ dữ liệu chuẩn của cả tiếng Anh, tiếng Việt đã chứng minh mô hình được đề xuất trong luận án có hiệu năng cao hơn đáng kể so với các mô hình trước đây. Bên cạnh đó, luận án cũng xây dựng bộ dữ liệu VLR999 đánh giá mô hình cho bài toán LER trong tiếng Việt, công bố bộ dữ liệu này cho công đồng nghiên cứu sử dụng.

Đối với bài toán phân tách các cặp từ theo quan hệ đồng nghĩa, trái nghĩa (Antonymy-Synonymy Classification), luận án đã đề xuất một mô hình mạng nơron (DVASNet) có khả năng khai thác các đặc trưng phân phối của từ trong kho ngữ liệu hay các vector nhúng từ. Ngoài ra, DVASNet còn nắm bắt được các thông tin về cấu trúc của từ. Kết quả thực nghiệm được đánh giá trên một số bộ dữ liệu chuẩn tiếng Việt đã chứng minh mô hình được đề xuất trong luận án có hiệu năng cao hơn từ 14% đến 17% theo độ đo $F1$ so với các mô hình trước đây.

Đối với bài toán đo lường độ tương tự ngữ nghĩa của cặp từ (Word Similarity Measurement), luận án đề xuất một mô hình cho đơn ngữ (mono-lingual Word Semantic Similarity) và một mô hình cho song ngữ (cross-lingual Word Semantic Similarity). Trong đó, mô hình đơn ngữ được đề xuất đã cải thiện hiện năng đo lường độ tương tự ngữ nghĩa của cặp từ tiếng Anh dựa trên thuật toán tìm đường đi ngắn nhất trên đồ thị. Đối với bài toán đo lường độ tương tự ngữ nghĩa của cặp từ song ngữ, luận án đã đề xuất một mô hình mạng nơron học không gian nhúng từ song ngữ Việt- Anh. Sử dụng không gian nhúng từ song ngữ để đo lường độ tương tự ngữ nghĩa cho các cặp từsong ngữ Việt-Anh. Bên cạnh đó, luận án cũng xây dựng bộ dữ liệu VSimLex-999, VESim-1000, công bố các bộ dữ liệu này cho cộng đồng nghiên cứu sử dụng.

1.4 Cấu trúc của luận án

Chương 2

KIẾN THỨC CƠ SỞ

2.1 Một số khái niệm cơ bản

2.2 Mô hình ngữ nghĩa phân phối

2.2.1 Khái niệm

Mô hình ngữ nghĩa phân phối (Distributional Semantic Models- DSMs) còn được biết đến là mô hình không gian từ (Word Space), mô hình không gian vector ngữ nghĩa (Vector Space Models-VSMs) hoặc phân phối tương tự (Distributional Similarity), là một mô hình biểu diễn nghĩa của từ bằng vector dựa trên phân phối của chúng trong kho ngữ liệu.

2.2.2 Lịch sử phát triển của DSMs

2.2.3 Cấu trúc DSMs

Một DSMs thường là một bộ gồm bảy thành phần $< T, C, R, W, M, d, S >$, bao gồm: T: tập hợp các đối tượng của không gian ngữ nghĩa, C: Ngữ cảnh (Context), R: Quan hệ giữa T và C, W: Lược đồ lượng giá trọng số, M: Không gian hình học hay ma trận đồng xuất hiện, d: Hàm giảm chiều ma trận $M \rightarrow M'$, S: độ đo tương tự áp dụng cho các vector trong ma trận M'.

2.3 Vector nhúng từ

Theo tiếp cận học máy và học sâu, các mô hình mạng nơron được sử dụng đòi hỏi các đối tượng cần xử lý phải được mã hóa bằng các vector đặc trưng. Theo đó, trong lĩnh vực NLP, các từ cần phải được biểu diễn bằng các vector ngữ nghĩa. Các phương pháp tạo ra vector biểu diễn cho từ được chia làm hai nhóm chính: tiếp cận dựa trên thống kê (Distributional Word vector Representaions) (Phần 2.2) và tiếp cận sử dụng mạng nơron học các biểu diễn vector dựa trên dự đoán sự đồng xuất hiện của các từ (Distributed Word vector Representaions), các mô hình này còn được gọi là mô hình vector nhúng từ (Word Embeddings Vector hay ngắn gọn hơn là Word Embeddings).

Chương 3

XÁC ĐỊNH QUAN HỆ BAO THUỘC DỰA TRÊN MÔ HÌNH NHÚNG TỪ CHUYÊN BIỆT

3.1 Đặt vấn đề

Bao thuộc (hypernymy) là một quan hệ cơ bản và quan trọng trong từ điển, cơ sở dữ liệu tri thứ tự vựng như WordNet và BabelNet. Quan hệ này có ứng dụng trong nhiều bài toán NLP như xây dựng cây ngữ nghĩa, phát hiện kế thừa văn bản, sinh văn bản... Quan hệ bao thuộc đang trở thành một chủ đề nghiên cứu được quan tâm trong NLP vì các ứng dụng của nó trong giải quyết các thách thức NLP như phát hiện ẩn dụ (metaphor detection). Luận án này đề xuất một mô hình phát hiện quan hệ bao thuộc dựa trên mô hình nhúng từ chuyên biệt.

3.1.1 Khái niệm

Bao thuộc (hypernymy) là một quan hệ ngữ nghĩa bất đối xứng (Asymmetric Semantic Relation) giữa một từ bao (hypernym) với một từ thuộc (hyponym). Ví dụ *động_vật* là một hypernym của *voi* hay *voi* là hyponym của *động_vật*. Tương tự, *hoa_hồng* là một từ bao của từ *hoa_hồng_bạch*, *xe_đạp_điện* là một từ thuộc của từ *xe_đạp*.

3.1.2 Tổng quan về bài toán xác định quan hệ hypernymy

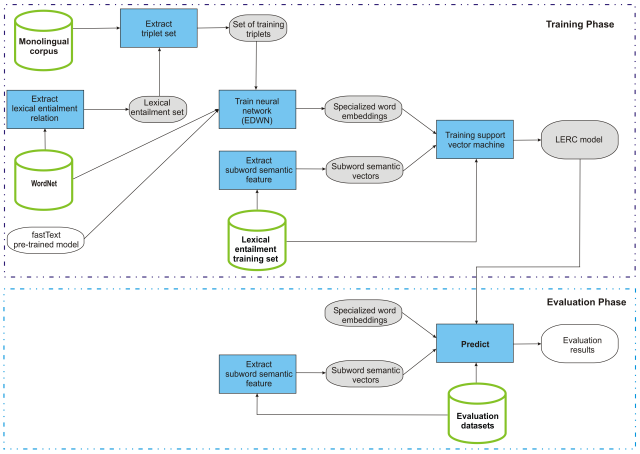
Các nghiên cứu về bài toán xác định quan hệ hypernymy (hypernymy recognition-HR) thường nhắm đề giải quyết bốn bài toán con (task) như sau: $f_{recognition} : (u, v) \rightarrow \{0, 1\}$, $f_{directionality} : (u, v) \rightarrow \{-1, 1\}$, $f_{determination} : (u, v) \rightarrow \{-1, 0, 1\}$.

3.2 Động lực của nghiên cứu này

3.3 Một số nghiên cứu liên quan

3.4 Mô hình LERC

Mô hình LERC khai thác thông tin từ cả kiến thức từ vựng và ngữ liệu để giải quyết vấn đề nhận dạng từ vựng. Hình 3.1 mô tả kiến trúc của mô hình này.



Hình 3.1: Tổng quan về mô hình LERC.

3.4.1 Mô hình DWN cải tiến

3.4.1.1 DWN with Attention

Chúng tôi ký hiệu $W_{V\times N}$ là một ma trận nhúng, trong đó V là kích thước của từ vựng, N là kích thước của vector nhúng. Mỗi từ ghép nghĩa, từ theo ngữ cảnh t được biểu diễn bằng một vectơ V -dimensional *one-hot* x_t, trong đó chỉ một trong số các phần tử V là 1 và các phần tử khác là 0 . v_t biểu thị N vectơ nhúng thứ nguyên của t, với v_t được tính theo công thức 3.1. Giả sử $x_p = 1$ và $x_{p'} = 0$ cho $p' \neq p$, v_t về cơ bản là hàng $p-th$ của ma trận nhúng.

$$v_t = x_t^T \cdot W \tag{3.1}$$

Đối với mỗi ngữ cảnh K -word, chúng tôi biểu thị $v_{context}$ là kết hợp có trọng số của K vectơ từ ngữ cảnh $v_{context}$ có thể được tính theo công thức 3.2, trong đó α_{c_i} là trọng số của từ theo ngữ cảnh c_i. Các trọng số này dưới dạng điểm chú ý được tính bằng các lược đồ trọng số chú ý khác nhau, được trình bày trong Phần 3.4.1.2.

$$v_{context} = \sum_{i=1}^{K} \alpha_{c_i} \cdot v_{c_i} \tag{3.2}$$

Biểu thị v_{hypo} là vectơ nhúng của từ ghép nghĩa và v_{comb} là một tổ hợp tuyến tính của $v_{context}$ và v_{hypo}, $\beta \in [0, 1]$ là một siêu tham số để cân bằng tác động của vectơ từ ghép và vectơ ngữ cảnh; do đó, v_{comb} được tính như sau:

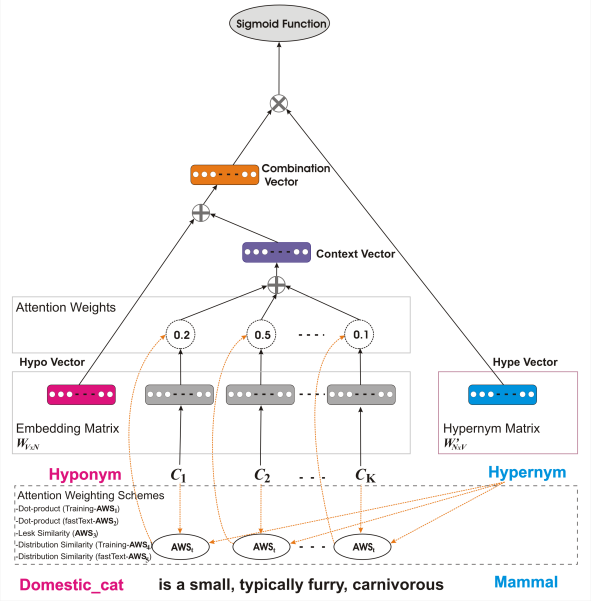
$$v_{comb} = \beta v_{hypo} + (1-\beta)\, v_{context} \tag{3.3}$$

Như trong Hình ref fig-DWN, có một ma trận trọng số khác $W'_{N\ times V}$ có vai trò là ma trận nhúng ẩn danh. Mỗi cột của W' là một vectơ chiều N v'_{hype} đại diện cho các siêu chữ. Sử dụng các trọng số này, chúng ta có thể tính điểm u_{hype} cho mỗi từ trong từ

vựng như sau:

$$u_{hype} = v'^{T}_{hype}.v_{comb} \tag{3.4}$$

Để đào tạo mô hình EDWN, chúng tôi sử dụng kỹ thuật lấy mẫu phủ định đã được



Hình 3.2: The architecture of the EDWN model.

sử dụng để đào tạo Word2Vec. Với mỗi triplet $<$ *hype*, *hypo*, các từ ngữ cảnh$>$, P mẫu phủ định $<$ $\overline{hype}$, *hypo*, các từ ngữ cảnh, $>$ được sinh ra, trong đó $\overline{hype}$ được sinh tuân theo phân phối unigram. Mô hình đề xuất được huấn luyện dựa trên mục tiêu làm cho v_{comb} và v'_{hype} tương tự nhau đối với các triplet xuất hiện trong corpus. Đồng thời, hàm mục tiêu muốn giữ cho v_{comb} khác với các vectơ ngẫu nhiên $v'_{\overline{hype}}$ trong các phủ định. Hàm sigmoid được sử dụng để chuẩn hóa sự giống nhau của hai vectơ tại lớp đầu ra của kiến trúc mạng nơron.

$$\sigma(v'^{T}_{hype/\overline{hype}} \cdot v_{comb}) \tag{3.5}$$

Mô hình EDWN được đào tạo bằng cách tuân theo giả thuyết rằng *các từ có các hyponym và ngữ cảnh tương tự phải được biểu diễn bằng các vector tương tự nhau.*

3.4.1.2 Lược đồ đo trọng số chú ý

Chúng tôi thực nghiệm với năm lược đồ đo trọng số của cơ chế chú ý (**A**ttention **W**eighting **S**cheme (AWS_i), $i \in [1..5]$), sử dụng các hàm đo lượng trọng số chú ý (Attention Function) khác nhau để tính toán điểm chú ý (Attention Score) bao gồm tích vô hướng của hai vector (dot product) và hai phép đo độ tương tự của từ, sử dụng ba nguồn kiến thức (knowledge sources) khác nhau bao gồm WordNet, mô hình EDWN đang được huấn luyện và mô hình fastText đã được huấn luyện trước. Bảng 3.1 mô tả tóm tắt các lược đồ đo trọng số chú ý này.

$$AWS_1\,(c_i, hype) = v^{T}_{c_i} \cdot v_{hype} \tag{3.6}$$

Bảng 3.1: Năm lược đồ tính trọng số chú ý.

Scheme	Attention Function	Attention Knowledge Source
AWS_1	Dot-product	Being trained
AWS_2	Dot-product	Pre-trained
AWS_3	Knowledge-based similarity measure	WordNet
AWS_4	Distribution similarity measure	Being trained
AWS_5	Distribution similarity measure	Pre-trained

$$AWS_3\left(c_i, hype\right) = Sim_{Lesk}\left(c_i, hype\right) = overlap\left(gloss\left(c_i\right), gloss\left(hype\right)\right) \tag{3.7}$$

$$AWS_4\left(c_i, hype\right) = \frac{v_{c_i} \cdot v_{hype}}{\|v_{c_i}\| \|v_{hype}\|} \tag{3.8}$$

Trong một trilet $< hype, hypo, cctngcnh >$, mỗi từ theo ngữ cảnh c_i có hệ số α_{c_i} tỷ lệ với điểm chú ý của c_i và $hype$, được tính bằng các Attention Function. Để tạo ra trọng số chú ý của các từ theo ngữ cảnh theo một phân phối xác suất, chúng tôi sử dụng hàm Softmax, theo đó α_{c_i} được tính như sau.

$$\alpha_{c_i} = softmax\left(AWS_l\left(c_i, hype\right)\right) = \frac{e^{AWS_l(c_i, hype)}}{\sum_{j=1}^{K} e^{AWS_l(c_j, hype)}} \tag{3.9}$$

Trong công thức 3.9, $l \in \{1, 2, 3, 4, 5\}$ có nghĩa là để tính trọng số chú ý, công thức này có thể sử dụng một trong năm lược đồ tính trọng số chú ý được trình bày ở trên. Lưu ý rằng $\sum_{i=1}^{K} \alpha_{c_i} = 1$, trong đó K là số từ ngữ cảnh.

3.4.2 Đặc trưng ngữ nghĩa mức dưới từ

Quan sát các cặp từ có quan hệ hypernymy của cả tiếng Việt và tiếng Anh, chúng tôi nhận thấy rằng có một mối tương quan ngữ nghĩa chặt chẽ giữa các thành phần của hai từ. Cụ thể, hai từ của một cặp hypernymy thường chia sẻ các thành phần chung hoặc các thành phần có quan hệ ngữ nghĩa với nhau. Hiện tượng này rõ ràng hơn trong các thuật ngữ của các lĩnh vực kỹ thuật.

3.4.2.1 Phân tích cấu trúc của các cặp hypernymy

Chúng tôi định nghĩa một mẫu ngữ nghĩa mức dưới từ (subword semantic pattern - SSP) là một cặp thành phần xuất hiện trong các cặp hypernymy mà giữa chúng tồn tại một quan hệ ngữ nghĩa như đồng nghĩa, trái nghĩa, bao thuộc.

3.4.2.2 Trích chọn đặc trưng

Chúng tôi đề xuất thuật toán 3.1 có thể nắm bắt các SSP của các cặp từ và nhúng chúng vào một vector (SSF).

Bảng 3.2: Thống kê số SSP và tỷ lệ các cặp từ xuất hiện SSP theo các quan hệ hypernymy, co-hyponymy, synonymy, antonymy, meronymy.

Quan hệ	Tiếng Việt		Tiếng Anh	
	#SSP	Tỷ lệ(%)	#SSP	Tỷ lệ(%)
Hypernymy	54	77.6	42	46.7
Co-hyponymy	31	27.7	29	15.4
Synonymy	23	21.5	11	5.3
Antonymy	16	12.3	8	2.6
Meronymy	12	11.7	7	3.1

Thuật toán 3.1 Trích chọn vector đặc trưng SSF của một cặp từ.

def Feature_Extraction(w_1, w_2)
a pair of words w_1-w_2 a pre-trained fastText model fT return a feature vector of w_1-w_2 $l_1 = \mathbb{S}(w_1)$; // generate a component set of w_1
$l_2 = \mathbb{S}(w_2)$; // generate a component set of w_2
$v_{SSF} = []$ **for** $i \leftarrow 0$ **to** $Length(l1)$ **do**
  for $j \leftarrow 0$ **to** $Length(l2)$ **do**
    $v_1 = fT[l_1[\text{i}]]$ $v_2 = fT[l_2[\text{j}]]$ $simScore = Cosine\_Similarity(v_1, v_2)$ $v_{SSF} = v_{SSF} \oplus [simScore]$; // $\oplus$ is the vector concatenation operator
  end
end
return v_{SSF}

3.4.3 Mô hình phân lớp quan hệ hypernymy có giám sát

$$v_{embeddings_{(x,y)}} = v_x \oplus v_y \oplus < v_x - v_y > \tag{3.10}$$

Để tạo ra vector đặc trưng đầu vào cho SVM, $v_{embeddings}$ được kết hợp với vectơ đặc trưng v_{SSF}. Bằng thực nghiệm, chúng tôi chọn phép nối vector làm toán tử kết hợp cho hai vector này. Toán tử nối ($\oplus$) được định nghĩa như sau:

$$v_{(x,y)} = v_{SSF_{(x,y)}} \oplus v_{embeddings_{(x,y)}} \tag{3.11}$$

3.5 Xây dựng bộ dữ liệu HR tiếng Việt

3.6 Thực nghiệm

Chúng tôi đã tiến hành các thực nghiệm để đánh giá hiệu năng của mô hình được đề xuất trong luận án đối với tiếng Việt và tiếng Anh, so sánh hiệu năng của mô hình này với một số mô hình tiêu biểu được công bố gần đây.

3.6.1 Bộ dữ liệu đánh giá

3.6.2 Huấn luyện các mô hình word embedding

3.6.3 Đánh giá mô hình

Thử nghiệm được thực hiện trên ba tác vụ LER với cài đặt tính năng 12, bao gồm *Word2Vec*(W2V), *GloVe*(GV), *fastText* (fT), *BERT*, *DWN*, textit EDWN, textit W2V $^+$, textit GV $^+$, *fT* $^+$, *BERT* $^+$, textit DWN $^+$, textit EDWN $^+$ (LERC). Ký hiệu ($^+$) biểu thị rằng đặc trưng SSF được sử dụng kết hợp với các vector word embedding (Công thức 3.11).

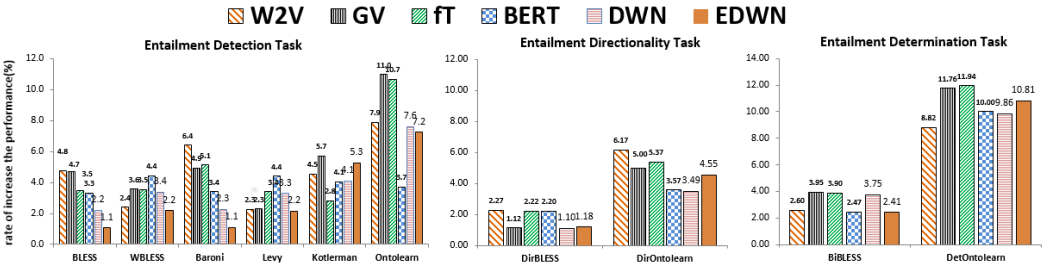
3.6.3.1 Đánh giá mô hình trên bộ dữ liệu tiếng Việt

Bảng 3.3: Đánh giá hiệu năng của các mô hình trên bộ dữ liệu tiếng Việt.

Model	**DtVLE999**(Δ%)	**DrVLE999**(Δ%)	**DetVLE999**(Δ%)
W2V	0.88	0.90	0.72
W2V$^+$	0.92(↑4.5)	0.93(↑3.3)	0.81(↑12.5)
GV	0.87	0.90	0. 72
GV$^+$	0.93(↑6.9)	0.94(↑4.4)	0.84(↑16.7)
fT	0.82	0.87	0.73
fT$^+$	0.89(↑8.5)	0.92(↑5.7)	0.83(↑13.7)
BERT	0.89	0.91	0.76
BERT$^+$	0.93(↑4.5)	0.94(↑3.3)	0.85(↑11.8)
DWN	0.86	0.89	0.77
DWN$^+$	0.89(↑3.5)	0.93(↑4.5)	0.86(↑11.7)
EDWN	0.94	0.95	0.82
LERC	**0.96**(↑2.1)	**0.97**(↑2.1)	**0.92**(↑12.2)

3.6.3.2 Đánh giá mô hình trên bộ dữ liệu tiếng Anh

Phần này trình bày các thực nghiệm với ba nhiệm vụ HR được tiến hành với tiếng Anh. Mức độ cải thiện hiệu năng của các mô hình trên các tập dữ liệu được minh họa trong Hình 3.3.



Hình 3.3: Tỷ lệ % của mức độ cải thiện hiệu năng của các mô hình khi được kết hợp với đặc trưng SSF, đánh giá trên các bộ dữ liệu tiếng Anh.

3.6.4 Phân tích đặc trưng SSF

3.7 Statistical Hypothesis Tests

3.8 Phân tích mô hình EDWN

3.9 Kết luận

Chương này của luận án trình bày về nghiên cứu *xác định quan hệ bao thuộc dựa trên mô hình nhúng từ chuyên biệt*. Nghiên cứu này có ba đóng góp quan trọng. Đầu tiên, đề xuất EDWN, là một mô hình nhúng từ chuyên biệt. Bằng cách thêm vào mô hình DWN ban đầu một lớp chú ý, mô hình EDWN có thể cung cấp các vectơ biểu diễn từ có chất lượng cao hơn cho các tác vụ HR. Thứ hai, giới thiệu đặc trưng SSF cũng như đề xuất phương pháp trích chọn đặc trưng này. SSF là một đặc trưng ngữ nghĩa hữu ích để nhận biết mối quan hệ hypernymy giữa các từ ghép và thuật ngữ kỹ thuật. Cuối cùng, luận án đề xuất mô hình LERC để giải quyết ba tác vụ HR bằng cách sử dụng kết hợp đặc trưng SSF và word embedding của mô hình EDWN. Thông qua nhiều thực nghiệm khác nhau trên bộ dữ liệu tiếng Anh và tiếng Việt, nghiên cứu đã cho thấy hiệu quả của phương pháp được đề xuất đối với bài toán HR, đặc biệt là đối với từ ghép, thuật ngữ kỹ thuật. Ngoài ra, đặc trưng ngữ nghĩa SSF có thể được sử dụng như một đặc trưng bổ sung hữu ích để nâng cao hiệu năng cho các mô hình bài toán nhận dạng quan hệ ngữ nghĩa khác. Cuối cùng, thông qua nghiên cứu này, luân án cung cấp sáu bộ dữ liệu đáng tin cậy được xây dựng với sự tham gia của các chuyên gia ngôn ngữ. Các bộ dữ liệu này được công bố để cộng đồng nghiên cứu tham chiếu, khai thác.

Chương 4

PHÂN TÁCH QUAN HỆ ĐỒNG NGHĨA – TRÁI NGHĨA DỰA TRÊN NGỮ CẢNH ĐỒNG HIỆN VÀ MẪU CẤU TRÚC TỪ

4.1 Đặt vấn đề

4.1.1 Khái niệm

Trái nghĩa (Antonymy) và đồng nghĩa (Synonymy) là những quan hệ paradigmatic. Những quan hệ này đóng vai trò quan trọng trong cấu trúc những cơ sở dữ liệu từ vừng tinh thần (mental lexicon knowledge). Trong đó, synonymy là quan hệ giữa hai từ tương đồng về ngữ nghĩa. Ngược lại, quan hệ antonymy được định nghĩa như sự đối ngịch về nghĩa giữa các từ.

4.1.2 Tổng quan về bài toán

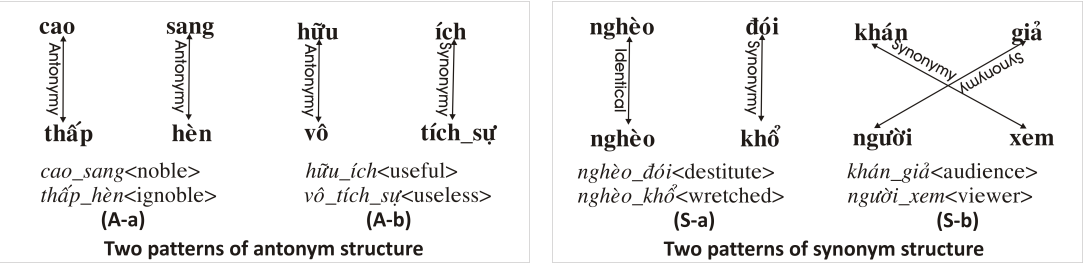
4.1.3 Một số nghiên cứu liên quan

4.1.4 Động lực của nghiên cứu này

Các mô hình ASC theo tiếp cận học sâu sử dụng mạng nơron khai thác thông tin Lexico-Syntactic của ngữ cảnh đồng hiện đạt hiệu năng vượt trội so với các cách tiếp cận khác. Tuy nhiên, các mô hình này yêu cầu kho ngữ liệu phải được phân tích cú pháp, yêu cầu này có thể kéo theo lỗi tích lũy gây ra bởi các công cụ phân tích cú pháp. Bên cạnh đó, sự phụ thuộc vào cấu trúc cú pháp có thể bỏ qua thông tin hữu ích như phủ định (negation), trạng từ (adverb), giới từ (preposition).

4.2 Các mẫu cấu trúc từ tiếng Việt (Vietnamese Word-Structure Patterns)

Qua khảo sát các cặp từ tiếng Việt theo các quan hệ ngữ nghĩa khác nhau (xem thêm Bảng 3.2 trong Chương 3), chúng tôi nhận thấy rằng *trong hai từ của một cặp từ ghép có quan hệ antonymy hoặc synonymy, các thành phần của từ này thường có quan hệ ngữ nghĩa với các thành phần của từ kia*, được gọi các mẫu cấu trúc từ (Word-Structure Pattern - WSP). Hình 4.1 là một minh họa trực quan của đặc trưng này. Trong Hình



Hình 4.1: Hình minh họa một số mẫu cấu trúc từ của các cặp trái nghĩa/đồng nghĩa.

4.1.(A-a), hai từ ghép trái nghĩa được cấu tạo bởi các từ đơn trái nghĩa. Trong Hình 4.1.(A-b), hai từ ghép trái nghĩa được cấu tạo bởi cặp từ đơn trái nghĩa và cặp từ đơn đồng nghĩa. Trong Hình 4.1.(S-a) hai từ ghép đồng nghĩa có chung một tiếng đầu và các thành phần còn lại của hai từ đồng nghĩa với nhau. Trong Hình 4.1.(S-b) một cặp từ ghép đồng nghĩa bao gồm hai cặp từ đơn đồng nghĩa trong đó có một cặp từ đồng nghĩa Hán Việt.

4.2.1 Mẫu cấu trúc từ trái nghĩa (Word Structure Patterns of Antonyms)

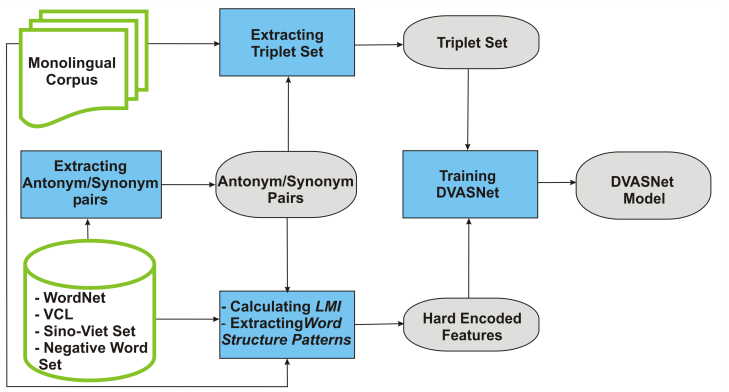
Khảo sát các cặp từ trái nghĩa tiếng Việt, chúng tôi xác định được 8 mẫu cấu trúc từ trái nghĩa (Word Structure Patterns of Antonyms, sau đây gọi tắt là Antonymy-Pattern) như sau: $\mathbf{u_{1}\_*}$ **-** $\mathbf{v_{1}\_*}$, $\mathbf{u_{1}\_u_{2}}$ **-** $\mathbf{v_{1}\_v_{2}}$, $\mathbf{u_{1}\_u_{2}}$ **-** $\mathbf{v_{1}\_v_{2}}$, $\mathbf{u_{1}\_x}$ **-** $\mathbf{v_{1}\_x}$, $\mathbf{u_{1}\_*}$ **-** $\mathbf{v}$, **u - nw_u**, $\mathbf{u_{1}\_*}$ **-** $\mathbf{nw\_u_{1}}$, $\mathbf{u_{1}\_u_{2}}$ **-** $\mathbf{v_{1}\_v_{2}}$.

4.2.2 Mẫu cấu trúc từ đồng nghĩa (Word Structure Patterns of Synonyms)

Khảo sát các cặp từ đồng nghĩa tiếng Việt, chúng tôi xác định được 7 mẫu cấu trúc từ đồng nghĩa (Word Structure Patterns of Synonyms, sau đây gọi tắt là Synonymy-Pattern) như sau: **x_* - x_***, **u - u_***, **u - *_u**, $\mathbf{u_{1}\_x}$ **-** $\mathbf{v_{1}\_x}$, $\mathbf{u_{1}\_*}$ **-** $\mathbf{v_{1}\_*}$, **x_y - y_x**, $\mathbf{u_{1}\_u_{2}}$ **-** $\mathbf{v_{1}\_v_{2}}$.

4.3 Đề xuất mô hình

Trong phần này của luận án, chúng tôi đề xuất một mô hình có thể khai thác các đặct trưng phong phú của tiếng Việt để phân loại từ trái nghĩa và từ đồng nghĩa. Đầu tiên chúng tôi trình bày về kiến trúc mạng nơron LSTM. LSTM hai chiều (biLSTM) là mô-đun quan trọng của DVASNet được sử dụng để mã hóa các ngữ cảnh đồng hiện trong một biểu diễn vectơ. Sau đó, chúng tôi mô tả kiến trúc của DVASNet kết hợp nhiều đặc trưng riêng của tiếng Việt để giải quyết vấn đề. Các đặc trưng được trích chọn không qua quá trình huấn luyện được gọi là đặc trưng tĩnh, là một vector được hình thành bởi ba đặc trưng thành phần được trích xuất trước khi huấn luyện mô hình. Tổng quan về mô hình DVASNet được trình bày trong Hình 4.2.



Hình 4.2: Tổng quan về mô hình phân tách cặp từ đồng nghĩa-trái nghĩa tiếng Việt.

Để huấn luyện mô hình, một tập các cặp từ trái nghĩa/đồng nghĩa được trích từ Mạng từ tiếng Việt và từ điển VCL. Sau đó, tập này được sử dụng để trích xuất một tập các bộ ba (triplet) từ một kho ngữ liệu. Một triplet là bộ ba thành phần gồm $<u, v, context>$ cùng xuất hiện trong một câu, với u và v là cặp từ đồng nghĩa hoặc trái nghĩa, *Context* là ngữ cảnh của cặp từ, bao gồm các từ trong câu (không bao gồm u và v) chứa cặp từ $u-v$. Ba đặc trưng của cặp từ trái nghĩa và từ đồng nghĩa bao gồm hệ số LMI, các WSP và hệ số tương tự ngữ nghĩa của cặp từ cũng được trích xuất dưới dạng các đặc trưng tĩnh hay mã hóa cứng. Cả hai loại đặc trưng được mã hóa mềm/cứng được khai thác để huấn luyện mô hình DVASNet như thể hiện trong Hình 4.3.

4.3.1 Kiến trúc mạng Long Short-Term Memory

Mạng nơron hồi quy (RNN) là một kiến trúc mạng thích hợp để mô hình hóa hóa dữ liệu có tính chất tuần tự vì nó cơ chế để lưu lại những trạng thái đã diễn ra. Ý tưởng nổi bật của RNN chính là kết nối các thông tin phía trước để dự đoán cho hiện tại.

$$i_i = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i) \tag{4.1}$$

$$f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f) \tag{4.2}$$

$$o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o) \tag{4.3}$$

$$g_t = tanh(W_g \cdot x_t + U_g \cdot h_{t-1} + b_g) \tag{4.4}$$

Tế bào nhớ hiện hành c_t là tổ hợp có trọng số của giá trị nhớ trước c_{t-1} và giá trị ứng viên g_t, các trọng số được tính theo giá trị cổng vào i_t và cổng ra f_t (Công thức 4.5).

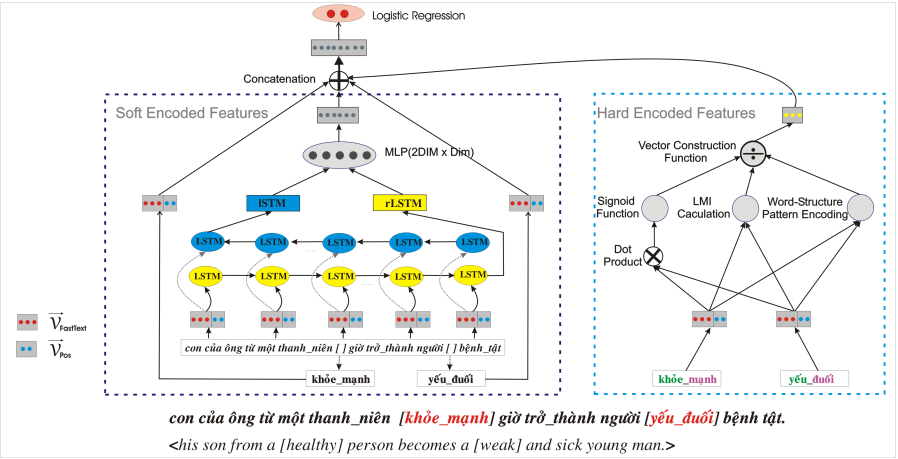
$$c_t = i_t \otimes g_t + f_t \otimes c_{t-1} \tag{4.5}$$

Đầu ra của một đơn vị LSTM (LSTM unit) là trạng thái ẩn h_t của của mạng hồi quy, h_t được tính theo công thức 4.6 như sau:

$$h_t = o_t \otimes tanh(c_t) \tag{4.6}$$

Trong Công thức 4.6, σ là ký hiệu của hàm Sigmoid, $\otimes$ là toán tử nhân Hadamard (Hadamard product hay element-wise product).

4.3.2 Mô hình DVASNet



Hình 4.3: Kiến trúc mạng nơron DVASNet cho bài toán phân lớp quan hệ đồng nghĩa-trái nghĩa.

$$\vec{v}_{biLSTM} = \vec{v}_{lLSTM(w_{1:n})} \oplus \vec{v}_{rLSTM(w_{n:1})} \tag{4.7}$$

Trong đó l/r biểu thị cho vector embedding *trái sang phải/từ phải sang trái* của ngữ cảnh được mã hóa bằng LSTM. Tiếp theo, một hàm kích hoạt tuyến tính được áp dụng cho vector biểu diễn ngữ cảnh như sau:

$$MLP(\vec{v}_{biLSTM}) = L_2(ReLU(L_1(\vec{v}_{biLSTM}))) \tag{4.8}$$

Với MLP là một mạng nơron truyền thẳng đa lớp (Multi Layer Perceptron), ReLU là hàm kích hoạt đơn vị tuyến tính (Rectified Linear Unit activation function), và $L_i(x) = W_i x + b_i$ là một lớp mạng nơron kết nối đầu đủ. Ký hiệu $\vec{v}_c$ là vector biểu diễn các từ ngữ cảnh, $\vec{v}_c$ được tính theo công thức 4.9 như sau:

$$\vec{v}_c = MLP(\vec{v}_{biLSTM}) \tag{4.9}$$

Ký hiệu $\vec{v}_{soft}$ là ghép nối của hai vector $\vec{v}_c$ và $\vec{v}_u$ and $\vec{v}_v$. $\vec{v}_{soft}$ được coi là vector mã hóa những đặc trưng "mềm" (hard encoded feature - những đặc có thể học được từ dữ liệu), $\vec{v}_{soft}$ được tính như sau:

$$\vec{v}_{soft} = \vec{v}_c \oplus \vec{v}_v \oplus \vec{v}_u \tag{4.10}$$

Ký hiệu $\vec{v}_{hard}$ là vector mã hóa các đặc trưng "cứng" (Hard Encoded Feature), là những đặc trưng được trích chọn trước pha huấn luyện mô hình từ tập các triplet. $\vec{v}_{hard}$ là vector $k-dimensions$ chiều được tạo bằng hàm sinh vector (Vector Construction

Bảng 4.1: Tỷ lệ xuất hiện mẫu cấu trúc từ của trái nghĩa/đồng nghĩa trong tiếng Việt.

Dataset	Số cặp	Số Patterns	Tỷ lệ (%)
ViCon	1,398	622	44.5
ViAS-1000	1,000	453	45.3
Antonym Pairs	24,347	15,126	62.1
Synonym Pairs	156,847	88,382	56.3

Function - VCF).

$$\vec{v}_{hard} = VCF(SimScore, LMI, WLP) \tag{4.11}$$

Cuối cùng, $\vec{v}_{triplet}$ là vector đặc trưng hợp nhất biểu diễn cho một trilet, là đặc trưng đầu vào cho tầng hồi quy logistic để phân tách cặp từ là đồng nghĩa với cặp từ trái nghĩa. $\vec{v}_{triplet}$ được tổ hợp như sau:

$$\vec{v}_{triplet} = \vec{v}_{hard} \oplus \vec{v}_{soft} \tag{4.12}$$

4.3.3 Các đặc trưng tĩnh

4.4 Xây dựng bộ dữ liệu ASC tiếng Việt

4.5 Thực nghiệm

4.5.1 Các mô hình cơ sở (Baseline Models)

4.5.2 Cài đặt thực nghiệm

4.5.2.1 Dữ liệu huấn luyện

4.5.2.2 Thiết lập tham số các tham số thực nghiệm

4.5.3 Kết quả thực nghiệm

4.5.3.1 Thống kê các mẫu cấu trúc từ

Để ước tính tỷ lệ các cặp từ đồng nghĩa/trái nghĩa phù hợp với các mẫu cấu trúc từ. Chúng tôi đã tiến hành phân tích tự động trên các tập dữ liệu bao gồm ViCon, ViAS-1000 (Bảng 4.1).

Bảng 4.2: Performance of the DVASNet model in comparison to the baseline models.

Model	ViCon			ViAS-1000		
	P	**R**	**F1**	**P**	**R**	**F1**
GloVe	0.78	0.63	0.70	**0.84**	0.45	0.59
fastText	0.76	0.63	0.69	0.79	0.44	0.57
Word2Vec	0.75	0.62	0.68	0.77	0.47	0.58
dLCE	**0.78**	0.67	0.72	0.80	0.55	0.65
Attract-Repel	0.76	0.67	0.71	0.75	0.47	0.58
DVASNet	0.72	0.80	0.76	0.75	0.70	0.72
DVASNet+	0.74	**0.82**	**0.78**	0.79	**0.73**	**0.76**

4.5.3.2 So sánh thông tin tương hỗ theo từng điểm giữa các cặp antonymy và synonymy

4.5.3.3 Phân tách cặp từ trái nghĩa với đồng nghĩa

Để đánh giá hiệu năng của mô hình, chúng tôi sử dụng độ đo F-Score (F1) làm thước đo đánh giá chính. F1 là trung bình điều hòa của độ đo Precision (P) và Recall (R). Kết quả thực nghiệm được trình bày trong Bảng 4.2, trong đó DVASNet là trường hợp mô hình chỉ sử dụng các đặc trưng được mã hóa mềm. DVASNet^{+} là trường hợp mô hình sử dụng cả đặc trưng được mã hóa mềm và cứng.

4.6 Kết luận

Chương này của luận án giới thiệu mô hình DVASNet, một mô hình mạng nơ-ron học sâu có thể khai thác hiệu quả các đặc trưng riêng của tiếng Việt cho bài toán ASC. Mô hình được đề xuất có thể sử dụng không chỉ thông tin từ vựng-cú pháp được thu thập từ các ngữ cảnh đồng xuất hiện của các cặp từ trong một kho ngữ liệu, mà còn cả các cấu trúc từ và các đặc trưng phân phối. Các mẫu cấu trúc từ lần đầu tiên được khai thác như một đặc trưng hữu ích của riêng tiếng Việt để nhận biết chính xác các quan hệ ngữ nghĩa. Ngoài ra, bộ dữ liệu ViAS-1000 cho bài toán ASC được xây dựng thỏa mãn nhiều tiêu chí khác nhau của tiếng Việt. Mô hình được đề xuất của chúng tôi đã vượt trội hơn đáng kể so với năm mô hình cơ sở từ 22% đến 25% về điểm R và từ 14% đến 17% của điểm F1.

Chương 5

ĐO LƯỜNG ĐỘ TƯƠNG TỰ NGỮ NGHĨA CỦA CẶP TỪ

5.1 Đặt vấn đề

Quan hệ tương tự là quan hệ Θ trên một tập không rỗng $\mathbb{X}$ thỏa mãn ba tính chất cơ bản như sau:

— Tính phản xạ (Reflexive): $x\ \Theta\ x$, $\forall\, x \in \mathbb{X}$.

— Tính đối xứng (Symmetric): Nếu $x\ \Theta\ y$ thì $y\ \Theta\ x$.

— Tính bắc cầu (Transitive): Nếu $x\ \Theta\ y$ và $y\ \Theta\ z$ thì $x\ \Theta\ z$.

Trong lĩnh vực ngôn ngữ học tính toán (Computational Linguistics), quan hệ tương tự về ngữ nghĩa (sau đây thuật ngữ "tương tự về ngữ nghĩa" được viết ngắn gọn là "tương tự") giữ các từ (Word Semantic Similarity hay Word Similarity) còn được gọi là sự tương đồng về đặc trưng phân loại của từ (Taxonomical Similarity) được dùng để chỉ các từ, khái niệm (Concept: *động_vật_hữu_nhũ*, *thực_vật_hạt_kín*, ...).

Các phương pháp WSM lượng giá mức độ giống nhau về nghĩa của hai từ, hay định lượng khoảng cách nhận thức giữa hai khái niệm với sự quan tâm về loại của chúng. Một mô hình WSM là một ánh xạ (f_{WSM}) từ tập các cặp từ sang tập các giá trị thực trong khoảng $[0, 1]$, như công thức 5.1 dưới đây.

$$f_{WSM} : (u, v) \rightarrow [0.0, 1.0] \tag{5.1}$$

5.2 Một số nghiên cứu liên quan

5.3 Đề xuất mô hình

Trong phần này luận án đề xuất hai mô hình WSM gồm ExtLeskSim và GraphSim.

5.3.1 Mô hình ExtLeskSim

Trong phần này của luận án, chúng tôi đề xuất một mở rộng thuật toán Lesk (**Ext**ended **Lesk** **Sim**ilarity - **ExtLeskSim**) để nó hoạt động hiệu quả hơn với đặc trưng của tiếng Việt và WordNet tiếng Việt, qua đó nâng cao hiệu suất của thuật toán

này cho bài toán WSM tiếng Việt.

$$ExtOverlap(w_1, w_2) = Overlap(w_1, w_2) + \sum_{u \in Hyponym(w_1)} Overlap(w_1, u) + \sum_{v \in Hyponym(w_2)} Overlap(w_2, v) + \sum_{\substack{x \in Hyponym(w_1) \\ y \in Hyponym(w_2)}} Overlap(x, y) \quad (5.2)$$

Khảo sát các Gloss được trích từ WordNet cho thấy rằng độ dài của chúng không đồng đều, ảnh hưởng đến độ chính xác của kết quả đo độ tương tự. Do đó, sử dụng một số độ đo tương tự giữa hai tập hợp như Szymkiewicz-Simpson, Cosine , Dice, Jaccard để chuẩn hóa kết quả của hàm Overlap theo độ dài của các Gloss có thể làm tăng độ chính xác của phép đo. Đối với mỗi độ đo tương tự của hai tập hợp như trên, chúng tôi xây dựng một độ đo tương tự giữa hai từ tương ứng. Trong các độ đo này, $EGloss(w)$ là ghép của *Gloss* của từ w và *Gloss* của các từ hyponym của nó.

$$ExtLeskSim_{Simpson}(w_1, w_2) = \frac{ExtOverlap(w_1, w_2)}{Min(Len(EGloss(w_1)), Len(EGloss(w_2)))} \quad (5.3)$$

$$ExtLeskSim_{Cosine}(w_1, w_2) = \frac{ExtOverlap(w_1, w_2)}{\sqrt{Len(EGloss(w_1)) \times Len(EGloss(w_2))}} \quad (5.4)$$

$$ExtLeskSim_{Dice}(w_1, w_2) = \frac{2 \times ExtOverlap(w_1, w_2)}{Len(EGloss(w_1)) + Len(EGloss(w_2))} \quad (5.5)$$

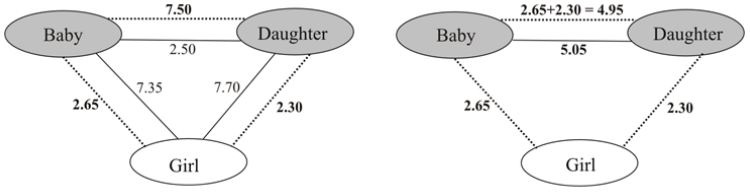
$$ExtLeskSim_{Jaccard}(w_1, w_2) = \frac{ExtOverlap(w_1, w_2)}{Len(EGloss(w_1)) + Len(EGloss(w_2)) - ExtOverlap(w_1, w_2)} \quad (5.6)$$

Để chuẩn hóa kết quả đo về miền giá trị từ 0 đến 10, chúng tôi sử dụng công thức 5.7, trong đó Sim_A là một trong các độ đo $Sim_{Simpson}$, Sim_{Cosine}, Sim_{Dice}, $Sim_{Jaccard}$.

$$Sim_A(w_1, w_2) = \frac{10 \times Sim_A(w_1, w_2)}{Max_{u,v \in VSimLex-999} Sim_A(u, v)} \quad (5.7)$$

5.3.2 Mô hình GraphSim

Phần này của luận án đề xuất một lược đồ cải tiến sử dụng cấu trúc đồ thị để nâng hiệu năng cho kỹ thuật WSM dựa trên WordNet (**Graph**-based **Sim**ilarity - **GraphSim**).



Hình 5.1: Một phần đồ thị tương tự của các cặp từ

Xuất phát từ ý tưởng có thể đo độ tương tự của một cặp từ dựa vào độ tương tự của chúng với những từ khác, Hình 5.1 trực quan hóa ý tưởng này, xác định độ tương tự của cặp từ *baby* và *daughter* thông qua từ trung gian *girl*, đường nối nét liền có trọng số giữa hai từ trong hình vẽ biểu thị cho độ tương tự của chúng, đường nối nét đứt có trọng số biểu thị cho khoảng cách giữa hai từ. Do tính đối ngẫu giữa độ đo tương tự và độ đo khoảng cách, chúng tôi đề xuất một lược đồ (GraphSim) nâng cao độ chính xác cho các kỹ thuật WSM, lược đồ GraphSim gồm năm bước như sau.

Bước 1: Xác định tập các từ phổ biến V.

Bước 2: Từ tập V, xây dựng tập các cặp từ phổ biến E.

Bước 3: Xây dựng đồ thị vô hướng có trọng số $G(V, E)$. Trong đó mỗi đỉnh của đồ thị là một từ thuộc V, mỗi cạnh của đồ thị tương ứng với một cặp từ thuộc E. Đồ thị G có ma trận trọng số W (Hình 5.1), trọng số của cạnh nối hai từ u, v được tính như sau:

$$W(u, v) = 10 - Similarity(u, v) \tag{5.8}$$

Bước 4: Sử dụng thuật toán Floyd-Warshall để tìm đường đi ngắn nhất giữa mọi cặp đỉnh của đồ thị G, tạo ra ma trận trọng số W' và ma trận lưu vết đường đi P. **Bước 5**: Sử dụng ma trận trong số W' để tính α:

$$\alpha = \frac{\sum_{u,v \in V} W'(u, v)}{\mid V \mid} \tag{5.9}$$

Bước 6: Tính độ tương tự của các cặp từ theo kỹ thuật m. Độ tương tự của từ u và v bằng hệ số α nếu chúng không liên thông trên đồ thị tương tự G. Ngược lại, nếu u liên thông với v, độ tương tự của hai từ được đo bằng tổng của độ tương tự tính theo kỹ thuật m với một lượng tỷ lệ thuận với độ tương tự cực đại và tỷ lệ nghịch với số đỉnh trung gian trên đường đi tối ưu giữa u và v. Trong đó $Length(u, v)$ là số đỉnh trung gian trên đường đi ngắn nhất từ u đến v.

$$Similarity(u, v) = \begin{cases} \alpha & \text{if } W'(u, v) \geq 10 \\ Similarity_m(u, v) + \frac{10 - W'(u,v)}{Length(u,v)} & \text{if } W'(u, v) < 10 \end{cases} \tag{5.10}$$

Trong đó độ dài của đường đi giữa u và v được tính như sau.

$$Length(u, v) = \begin{cases} 1 & \text{if } p[u, v] = 0 \\ Length(u, p[u, v] + Legth(p[u, v], v)) & \text{if } p[u, v] \neq 0 \end{cases} \tag{5.11}$$

Chúng tôi lựa chọn một tập các cặp từ được sử dụng phổ biến trong tiếng Anh để xây dựng một đồ thị ngữ nghĩa có trọng số không âm. Để tìm đường đi tối ưu trên đồ thị, chúng tôi sử dụng thuật toán Floyd-Warshall [1]. Thuật toán này được Robert Floyd đề xuất năm 1962 cho bài toán xác định đường đi ngắn nhất giữa mọi cặp đỉnh của đồ thị (Thuật toán 5.1).

1. https://en.wikipedia.org/wiki/Floyd%E2%80%93Warshall_algorithm

Thuật toán 5.1 Thuật toán tìm đường đi ngắn nhất giữa mọi cặp đỉnh trên đồ thị.

def Floyd-Warshall()

ma trận trọng số W; ma trận trọng số W', ma trận lưu vết đường đi P **for** $k \leftarrow 0$ **to** n **do**

```
    for i ← 0 to n do
        for j ← 0 to n do
            if W[i,j] > W[i,k]+W[k,j] then
                W[i,j] = W[i,k]+W[k,j]  P[i,j] = k
            else
        end
    end
end
return W'
```

5.4 Xây dựng bộ dữ liệu tiếng Việt

5.4.1 Dịch bộ dữ liệu SimLex-999 sang tiếng Việt

5.4.2 Đánh giá độ tương tự của cặp từ

5.4.3 Một số thống kê trên bộ dữ liệu

5.5 Thực nghiệm

Phần này trình bày các thực nghiệm với các bộ dữ liệu tiếng Anh **Simlex-999**, và tiếng Việt **VSimlex-999**, **ViSim-400**.

5.5.1 Thực nghiệm với mô hình ExtLeskSim

Bảng 5.1: Kết quả đánh giá hiệu năng theo độ tương quan Pearson của các mô hình trên bộ dữ liệu VSimLex-999 và SimLex-999.

Mô hình	**VSimLex-999**				**SimLex-999**
	Total	A	N	V	Total
CBOW	0.52	0.47	0.52	0.55	-
SG	0.53	0.50	0.51	0.62	0.44
Trung bình	0.53	0.49	0.52	0.59	0.44
WuP	0.27	-	0.29	0.14	0.32
Path	0.40	-	0.46	0.25	0.45
LC	0.31	-	0.41	0.12	0.29
Resnik	0.28	-	0.30	0.13	0.35
JC	0.20	-	0.24	0.07	0.20
Lin	0.37	-	0.41	0.22	0.39
Trung bình	0.28	-	0.32	0.14	0.33
Lesk	0.42	0.50	0.42	0.50	0.35

Để đánh giá hiệu năng của các kỹ thuật ExtLeskSim, Gloss của các từ thuộc VSimLex-999 và các từ là hyponym của chúng được trích từ WordNet. Kết quả thực nghiệm được trình bày trong Bảng 5.2.

Bảng 5.2: Hiệu năng theo độ đo hệ số tương quan Pearson của các kỹ thuật ExtLeskSim.

Mô hình	Trung bình	Tính từ	Danh từ	Động từ
Lesk	0.41	0.47	0.40	0.47
ExtLeskSim$_{Cosine}$	0.43	0.46	0.44	0.49
ExtLeskSim$_{Dice}$	**0.49**	0.56	0.44	0.58
ExtLeskSim$_{Jaccard}$	0.48	0.57	0.40	0.59
ExtLeskSim$_{Simpson}$	**0.49**	0.59	0.43	0.57

5.5.2 Thực nghiệm với mô hình GraphSim

Để đánh giá hiệu quả của lược đồ cải tiến GraphSim, chúng tôi thực nghiệm các kỹ thuật WSM gốc (Wu-Palmer, Leacock-Chodorow, Resnik, Jiang-Conrath, Lin) và kỹ thuật cải tiến có sử dụng GraphSim. Kết quả thực nghiệm được trình bày trong Bảng 5.3.

Bảng 5.3: Kết quả thực nghiệm lược đồ cải tiến.

Mô hình	Mô hình gốc	GraphSim
WuP	0.32	0.43
Path	0.45	0.48
LCh	0.29	0.39
Res	0.35	0.44
JC	0.20	0.32
Lin	0.39	0.45

5.6 Kết luận

Chương này của luận án đã đề xuất kỹ thuật ExtLeskSim và GraphSim. Kỹ thuật ExtLeskSim đo lường độ tương tự ngữ nghĩa của cặp từ sử dụng thông tin định nghĩa của từ. Kết quả thực nghiệm đã cho thấy ExtLeskSim đạt hiệu năng cao đối với tiếng Việt. Kỹ thuật GraphSim đã cải thiện hiệu năng đo lường độ tương tự ngữ nghĩa của các kỹ thuật WSM dựa trên WordNet. Kết quả thực nghiệm trên bộ dữ liệu tiếng Anh cho thấy GraphSim đã nâng cao đáng kể hiệu năng cho các kỹ thuật WSM được áp dụng.

Chương 6

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1 Các đóng góp của luận án

Tự động xác định quan hệ ngữ nghĩa của từ là một trong những bài toán quan trọng và khó khăn nhất của NLP. Luận án này đã khai thác tiếp cận ngữ nghĩa phân phối để định lượng cũng như định tính một số quan hệ ngữ nghĩa từ vựng cơ bản gồm similarity, synonymy, antonymy, hypernymy (Bảng 1.1), cho cả tiếng Việt và tiếng Anh. Thực hiện đề tài *Tự động xác định quan hệ ngữ nghĩa của từ dựa trên học máy thống kê*, luận án đã có những đóng góp chính như sau:

Đối với bài toán Hypernymy Recognition:

1. Đề xuất một lược đồ trích chọn những đặc trưng ngữ nghĩa mức dưới từ (Subword Semantic Feature). Lược đồ được đề xuất không những mã hóa được quan hệ ngữ nghĩa của các thành phần của cặp từ mà còn nắm bắt được cả thông tin vị trí của chúng trong các vector đặc trưng ngữ nghĩa dưới từ.

2. Đề xuất một mô hình mạng nơron học sâu EDWN học các vector biểu diễn từ chuyên biệt (specialized word embedding vector). Các vector của EDWN phù hợp cho bài toán Hypernymy Recognition hơn Word2vec, fastText, GloVe.

3. Đề xuất mô hình LERC, mô hình này đã sử dụng với đặc trưng đầu vào được kết hợp từ vector nhúng từ và vector đặc trưng ngữ nghĩa dưới từ. Kết quả thực nghiệm được đánh giá trên một số bộ dữ liệu chuẩn của cả tiếng Anh, tiếng Việt đã chứng minh mô hình được đề xuất trong luận án có hiệu năng cao hơn đáng kể so với các mô hình trước đây.

4. Xây dựng bộ dữ liệu VLR999, đây là bộ dữ liệu đánh giá mô hình cho bài toán Hypernymy Recognition tiếng Việt đầu tiên được công bố cho công động nghiên cứu sử dụng.

Những đóng góp đối bài toán Hypernymy Recognition được trình bày trong các công bố (9), (11) trong mục danh sách công trình khoa học.

Đối với bài toán Antonymy-Synonymy Classification:

5. Đề xuất các mẫu cấu trúc từ tiếng Việt (Vietnamese Word-Structure Patterns). Các mẫu cấu trúc từ đã được chứng minh là những đặc trưng hữu ích giúp nâng cao độ chính xác của phân lớp các cặp từ đồng nghĩa và trái nghĩa.

6. Đề xuất một mô hình mạng nơron học sâu DVASNet, mô hình này đã khai thác ngữ cảnh đồng hiện của các cặp từ mà không cần dựa vào cây phân tích cú pháp. Mô hình DVASNet cũng kết hợp các các đặc trưng tĩnh như mẫu cấu trúc từ, độ tương tự ngữ nghĩa, thông tin tương hỗ giữa hai từ với đặc trưng thông tin ngữ cảnh đồng hiện để giải quyết hiệu quả bài toán ASC cho tiếng Việt. Kết quả thực nghiệm được đánh giá trên một số bộ dữ liệu chuẩn tiếng Việt đã chứng minh mô hình DVASNet có hiệu năng cao hơn từ 14% đến 17% theo độ đo $F1$ so với các mô hình được đề xuất trước đây.

Những đóng góp được trình bày trong các công bố (6), (7), và (8) trong mục danh sách công trình khoa học.

Đối với bài toán Word Similarity Measurement:

7. Đề xuất kỹ thuật ExtLeskSim đo lường độ tương tự ngữ nghĩa của cặp từ sử dụng thông tin định nghĩa của từ. Kết quả thực nghiệm đã cho thấy ExtLeskSim đạt hiệu năng cao đối với tiếng Việt.

8. Đề xuất lược đồ GraphSim đã cải thiện hiện năng đo lường độ tương tự ngữ nghĩa của các kỹ thuật WSM dựa trên WordNet. Kết quả thực nghiệm trên bộ dữ liệu tiếng Anh cho thấy GraphSim đã nâng cao đáng kể hiệu năng cho các kỹ thuật WSM được áp dụng.

9. Đối với bài toán đo lường độ tương tự ngữ nghĩa của cặp từ song ngữ, luận án đã đề xuất một mô hình mạng nơron học mô hình word embedding song ngữ Việt- Anh. Sử dụng mô hình word embedding song ngữ đãn học được để đo lường độ tương tự ngữ nghĩa cho các cặp từ song ngữ Việt-Anh.

10. Xây dựng bộ dữ liệu VSimLex-999, VESim-1000, công bố các bộ dữ liệu này cho công động nghiên cứu sử dụng.

Những đóng góp được trình bày trong các công bố (1), (10), (13) và (14) trong mục danh sách công trình khoa học.

6.2 Hướng phát triển

Luận án đã đề xuất một số mô hình hiệu quả cho bài toán xác định quan hệ ngữ nghĩa của từ vựng. Trên cơ sở những nghiên cứu đã tiến hành, một số hướng nghiên cứu tiếp theo của luận án có thể được tiến hành gồm: đề xuất mô hình học vector biểu diễn từ với đặc trưng ngữ nghĩa phu hợp cho bài toán LE, nghiên cứu sử dụng các mô hình học sâu để khai thác nhiều đặc trưng ngữ nghĩa của từ giúp tăng hiệu năng cho mô hình ASC, nNghiên cứu đo lường độ tương tự của cặp từ theo ngữ cảnh.

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN TỚI LUẬN ÁN

(1). Van-Tan Bui and Phuong-Thai Nguyen. *WEWD: A Combined Approach for Measuring Cross-lingual Semantic Word Similarity Based on Word Embeddings and Word Definitions*. The 2021 RIVF International conference on computing and communication technologies, 2021. (Scopus, DBLP)

(2). Van-Tan Bui and Phuong-Thai Nguyen. *Measuring semantic similarity of vietnamese sentences based on lexical similarity and distribution semantic similarity*. Advances in Intelligent Systems and Computing of Springer-Verlag 2021 (AISC Series). (SCOPUS, IF=0.9).

(3). Van-Tan Bui and Phuong-Thai Nguyen, Van-Lam Pham. *Combining Specialized Word Embeddings and Subword Semantic Features for Lexical Entailment Recognition*. ACM Transactions on Asian and Low-Resource Language Information Processing, 2021, (Submitted).

(4). Hong-Viet Tran, Van-Tan Bui, Dinh-Tien Do, Van-Vinh Nguyen. *Combining PhoBERT and SentiWordNet for Vietnamese Sentiment Analysis*. The 13th International Conference on Knowledge and Systems Engineering (KSE), 2021. (Scopus, DBLP).

(5). Van-Tan Bui and Phuong-Thai Nguyen. *A Combined Model Measuring Vietnamese Sentences Similarity Based on Lexical Correlation and Word Embeddings*. Fundamental and Applied IT Research (Fair), 2021, (Submitted).

(6). Van-Tan Bui , Phuong-Thai Nguyen and Khac-Quy Dinh. *Vietnamese Antonyms Detection Based on Specialized Word Embeddings using Semantic Knowledge and Distributional Information*. The 12th International Conference on Knowledge and Systems Engineering (KSE), 2020. (Scopus, DBLP)

(7). Van-Tan Bui, Phuong-Thai Nguyen, Van-Lam Pham and Thanh-Quy Ngo. *A Neural Network Model for Efficient Antonymy-Synonymy Classification by Exploiting Co-occurrence Contexts and Word-Structure Patterns*. International Journal of Intelligent Engineering and Systems, Vol.13, No.1, 2020. (SCIE Journal, IF=1.9)

(8). Bui Van Tan, Nguyen Phuong Thai, Pham Van Lam and Dinh Khac Quy. *Antonyms-Synonyms Discrimination Based On Exploiting Rich Vietnamese Features*. 16th International Conference of the Pacific Association for Computational Linguistics, 2019. (Scopus, DBLP)

(9). Bui Van Tan and Nguyen Phuong Thai. *Enhancing Performance of Lexical Entailment Recognition for Vietnamese based on Exploiting Lexical Structure Features*. The 10th International Conference on Knowledge and Systems Engineering (KSE), 2018. (Scopus, DBLP)

(10). Bui Van Tan, Nguyen Phuong Thai and Dinh Khac Quy. *Cross-lingual Semantic Similarity via Cross-Lingual Embeddings*. Fundamental and Applied IT Research Conference (Fair), 2018.

(11). Bui Van Tan, Nguyen Phuong Thai and Pham Van Lam. *Hypernymy Detection for Vietnamese Using Dynamic Weighting Neural Network*. 19th International Conference on Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science. (Scopus,IF=1.8)

(12). Minh-Thuan Nguyen, Van-Tan Bui, Huy-Hien Vu, Phuong-Thai Nguyen, Chi-Mai Luong. *Enhancing the quality of Phrase-table in Statistical Machine Translation for Less-Common and Low-Resource Languages*. International Conference on Asian Language Processing, 2018. (Scopus, DBLP)

(13). Bui Van Tan, Nguyen Phuong Thai and Pham Van Lam. *Construction of a Word Similarity Dataset and Avaluation of Word Similarity Techniques for Vietnamese*. The 9th International Conference on Knowledge and Systems Engineering (KSE), 2017. (Scopus, DBLP)

(14). Bui Van Tan, Nguyen Phuong Thai and Nguyen Minh Thuan. *Enhancement of Measurement Efficency for Semantic Similarrity based on WordNet*. Fundamental and Applied IT Research (Fair), 2017.