ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Hà Mỹ Linh

NGHIÊN CỨU CÁC PHƯƠNG PHÁP BIỂU DIỄN VÀ PHÁT TRIỂN NGỮ LIỆU, CÔNG CỤ CHO PHÂN TÍCH CÚ PHÁP VÀ NGỮ NGHĨA TIẾNG VIỆT

Ngành : **Cơ sở toán học cho Tin học**
Mã số : **9460117.02**

TÓM TẮT LUẬN ÁN TIẾN SĨ TOÁN TIN

Hà Nội - 2025

Công trình được hoàn thành tại:

Trường Đại học Khoa học Tự nhiên
Đại học Quốc gia Hà Nội

Người hướng dẫn khoa học:
1. GS.TS. Nguyễn Lê Minh
2. TS. Nguyễn Thị Minh Huyền

Phản biện 1: PGS.TS Nguyễn Lưu Thuỳ Ngân.
Trường Đại học Công nghệ Thông tin, Đại học Quốc gia TPHCM.
Phản biện 2: PGS.TS Thân Quang Khoát.
Trường Công nghệ Thông tin và Truyền thông, Đại học Bách khoa Hà Nội.
Phản biện 3: PGS. TS Trịnh Cẩm Lan.
Trường Đại học Khoa học Xã hội và Nhân văn, Đại học Quốc gia Hà Nội.

Luận án được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ họp tại

Trường Đại học Khoa học Tự nhiên, ĐHQGHN

Vào hồi 14 giờ 30 phút, ngày 28 tháng 06 năm 2025

Có thể tìm hiểu luận án tại thư viện:

1. **Thư viện Quốc gia Việt Nam**
2. **Trung tâm Thư viện và Tri thức số, Đại học Quốc gia Hà Nội**

DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ CỦA LUẬN ÁN

[P1] **Ha My Linh**, Nguyen Thi Minh Huyen, "*A Case Study on Meaning Representation for Vietnamese*", Proceedings of the First International Workshop on Designing Meaning Representations, pages 148-153, 2019, Italy, (*ISBN 978-1-950737-45-1*).

[P2] **Ha My Linh**, Nguyen Thi Minh Huyen, Vu Xuan Luong, Nguyen Thi Luong, Phan Thi Hue, Le Van Cuong, "*VLSP 2020 shared task: Universal dependency parsing for Vietnamese*", Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing, pages 77-83, 2020, Vietnam.

[P3] **Ha My Linh**, Le Van Cuong, Nguyen Thi Minh Huyen, "*Construction of a VerbNet style Lexicon for Vietnamese*", Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation, pages 84-91, 2020, Vietnam, (*ISSN 2619-7782*).

[P4] **Ha My Linh**, Do Duy Dao, Nguyen Thi Minh Huyen, Tran Thu Trang, "*Using rules for building Vietnamese AMR-based corpus*", Một số vấn đề chọn lọc của công nghệ thông tin và truyền thông lần thứ XXIV, pages: 547-552, 2021, Vietnam.

[P5] Ishan Jindal, ALexandre Rademaker, Michał ULewicz, **Ha My Linh**, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu, and Yunyao Li, "*Universal Proposition Bank 2.0*". In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 1700 – 1711, 2022, MarseilLe, France, European Language Resources Association, (*ISBN: 979-10-95546-72-6*).

[P6] **Ha My Linh**, Do Duy Dao, Nguyen Thi Minh Huyen, Ngo The Quyen, Doan Xuan Dung, "*VLSP 2021 - NER Challenge: Named Entity Recognition for Vietnamese*", VNU Journal of Science: Computer Science and Communication Engineering, pages 87 - 97, Vol 38.1, 2022, (*VNU Journal of Science ISSN: 0866-8612*).

[P7] The Quyen Ngo, Thi Anh Phuong Nguyen, **Ha My Linh**, Thi Minh Huyen Nguyen, Phuong Le-Hong, "*Improving Multi-label Classification of Similar Languages by Semantics-Aware Word Embeddings*", In Eleventh Workshop on NLP for Similar Languages, Varieties and DiaLects (VarDial 2024), pages 235-240, Mexico, Association for Computational Linguistics, (*ISBN 979-8-89176-104-9*).

[P8] **Ha My Linh**, Thi Minh Huyen Nguyen, The Quyen Ngo, Tuan Thanh Le, Tran Thai Dang, Viet Hoang Ngo, Xuan Dung Doan, Thi Luong Nguyen, Van Cuong Le, Thi Hue Phan, Xuan Luong Vu, "*VLSP 2022 Challenge: Vietnamese Constituency Parsing*", **accepted** to Journal of Computer Science and Cybernetics, 2025, (*ISSN 2815-5939*).

[P9] **Ha My Linh**, Pham Thi Duc, Le Ngoc Toan, Thi Minh Huyen Nguyen, "*An Investigation of ISO-TimeML Applied to Vietnamese*", Proceedings of the 38th Pacific Asia Conference on Language, Information, and Computation, PACLIC 38 (2024), Tokyo, Japan, pages 1387 - 1394.

GIỚI THIỆU

Xử lí ngôn ngữ tự nhiên (*Natural Language Processing - NLP*) đã thu hút nhiều sự quan tâm của các nhóm nghiên cứu trên thế giới ngay từ khi máy tính điện tử ra đời. Nhiều phương pháp đã được phát triển để giải quyết các bài toán phân tích cú pháp và ngữ nghĩa, từ các cách tiếp cận dựa trên luật, cho đến các kỹ thuật học máy, đặc biệt là các mô hình hiện đại sử dụng học sâu và gần đây nhất là sự phát triển của các mô hình ngôn ngữ lớn. Trong hầu hết những cách tiếp cận để giải quyết bài toán cú pháp và ngữ nghĩa, việc xây dựng tài nguyên từ vựng và các kho văn bản có chú giải ngôn ngữ là vô cùng cần thiết, có giá trị cao và tính ổn định lâu dài, đều được dựa trên quy trình chuẩn hóa.

Động lực nghiên cứu

Các tài nguyên từ vựng chứa các thông tin về hình thái từ, các khung cú pháp, ngữ nghĩa, các ràng buộc và mối quan hệ giữa các thành phần câu với từ vựng đó. Một số kho từ vựng nổi bật như: WordNet [41, 86], FrameNet [13]; hay VerbNet [68]. Thuật ngữ "ngân hàng cây" (*treebank*) chỉ các văn bản được chú giải thông tin từ loại, cú pháp chi tiết, tạo cơ sở cho các bài toán phân tích từ, cú pháp và ngữ nghĩa. Về ngữ nghĩa, các mô hình biểu diễn và kho ngữ liệu có chú giải ngữ nghĩa cũng đã và đang được các nhóm nghiên cứu quan tâm và phát triển để có thể hình thức hóa nghĩa của từ, câu và đoạn văn. Các kho ngữ liệu có chú giải ngữ nghĩa tiêu biểu gồm Propbank [67], AMR [14] hay các ngân hàng ngữ nghĩa khác như Groningen - GMB [58], UCCA [7], ...

Đối với tiếng Việt, việc phát triển kho từ vựng và các kho ngữ liệu có chú giải ngữ pháp, ngữ nghĩa cũng đã được quan tâm từ nhiều năm trước. Tài nguyên từ vựng đầu tiên được xây dựng từ năm 2006 là Từ điển tiếng Việt cho máy tính [94] (*Vietnamese Computational Lexicon* – VCL), kho từ vựng WordNet tiếng Việt [1], ngân hàng cây cho tiếng Việt được xây dựng từ năm 2009 là Viettreebank [101], các ngân hàng cây cú pháp phụ thuộc với những tập nhãn phụ thuộc riêng của từng nhóm [30, 65, 92]. Đối với bài toán phân tích ngữ nghĩa, một tập nhãn vai nghĩa cùng kho ngữ liệu gồm 5,640 câu đã được xây dựng [2].

Tính cấp thiết của đề tài

Các nghiên cứu về ngữ pháp và ngữ nghĩa tiếng Việt đã đạt được những thành tựu đáng kể, tuy nhiên vẫn còn nhiều thách thức như: các bộ nhãn thành phần

và phụ thuộc của các nhóm xây dựng riêng, thiếu sự thống nhất và chuẩn hóa, nhiều nghiên cứu chưa cung cấp thông tin chi tiết về quá trình xây dựng bộ nhãn, quy trình gán nhãn dữ liệu. Về ngữ nghĩa, hiện tại chưa có một mô hình hay kho ngữ liệu nào được gán nhãn ngữ nghĩa đầy đủ và toàn diện cho tiếng Việt. Kho ngữ liệu gán nhãn vai nghĩa đã xây dựng chưa được chuẩn hóa và liên kết chặt chẽ với các kho ngữ liệu khác, hiệu quả của các mô hình gán nhãn vai nghĩa cũng còn khá hạn chế.

Mục tiêu nghiên cứu

Từ tình hình trên, luận án này đặt mục tiêu nghiên cứu phát triển ngữ liệu cùng các sơ đồ chú giải, bao gồm kho từ vựng cũng như các kho ngữ liệu có chú giải cú pháp, ngữ nghĩa, tuân theo các mô hình chuẩn hoá tài nguyên ngôn ngữ trên thế giới. Song song với xây dựng ngữ liệu, luận án cũng đánh giá, phát triển các công cụ phân tích cú pháp và ngữ nghĩa tiếng Việt, hỗ trợ qua lại công việc xây dựng ngữ liệu.

Cụ thể, những công việc được thực hiện trong luận án được mô tả chi tiết trong Hình 1.

Hình 1: *Mục tiêu của luận án.*

Phạm vi nghiên cứu

Để đạt được các mục tiêu trên, luận án sẽ giải quyết các bài toán sau:

- Phân tích cú pháp: Xây dựng tập nhãn cú pháp, kho ngữ liệu và phát triển các công cụ phân tích cú pháp thành phần, cú pháp phụ thuộc.

- Phân tích ngữ nghĩa câu: Xây dựng tập nhãn vai nghĩa, kho ngữ liệu, xây dựng mô hình biểu diễn ngữ nghĩa cho văn bản tiếng Việt, thử nghiệm một số mô hình phân tích ngữ nghĩa cho tiếng Việt.
- Phân tích ngữ nghĩa từ vựng: Nghiên cứu và thiết kế, xây dựng mạng động từ (viVerbnet) cho tiếng Việt.

Đóng góp của luận án

Luận án đã có những đóng góp cơ bản về hai hướng chính:

- Xây dựng các lược đồ chú giải và kho ngữ liệu: Cú pháp phụ thuộc (xây dựng lại tập nhãn cú pháp phụ thuộc, kho ngữ liệu gồm hơn 9,000 câu), cú pháp thành phần (rà soát, cập nhật và chuẩn hoá các nhãn cú pháp thành phần và tài liệu hướng dẫn gán nhãn, xây dựng kho ngữ liệu gồm hơn 9,000 câu), xây dựng kho ngữ liệu gán nhãn vai nghĩa cho tiếng Việt (gồm 2,570 câu), xây dựng mô hình và hướng dẫn gán nhãn ngữ nghĩa cho tiếng Việt dựa vào mô hình ngữ nghĩa trừu tượng của tiếng Anh (AMR) và các vai nghĩa LIRICS [98]. Kho ngữ liệu tiếng Việt gồm có 1,570 câu đã được xây dựng.
- Về phương pháp và mô hình: Luận án đã thử nghiệm một số mô hình phân tích cú pháp phụ thuộc, khảo sát các phương pháp phân tích cú pháp thành phần, và thảo luận về kết quả đạt được. Xây dựng mô hình biểu diễn và chú giải ngữ nghĩa cho tiếng Việt, thử nghiệm các mô hình ngôn ngữ lớn để gán nhãn vai nghĩa và phân tích ngữ nghĩa cho văn bản tiếng Việt, đánh giá và phân tích kết quả đạt được. Thu thập dữ liệu, trích rút ngữ cảnh và các động từ trong tiếng Việt. Sau đó, sử dụng các thuật toán phân cụm các động từ tiếng Việt. Thiết kế mạng động từ tiếng Việt (viVerbNet) dựa vào VerbNet tiếng Anh gồm có 5 thành phần chính.

Cấu trúc của luận án

Luận án được tổ chức như sau:

- Chương 1: Trình bày các kiến thức cơ sở.
- Chương 2: Mô tả chi tiết về việc xây dựng tài nguyên và công cụ phân tích cú pháp tiếng Việt.
- Chương 3: Xây dựng tài nguyên và công cụ phân tích ngữ nghĩa tiếng Việt.
- Chương 4: Trình bày về việc xây dựng mạng động từ tiếng Việt (viVerbNet).
- Phần kết luận: Tóm tắt một số kết quả đạt được và hướng phát triển trong tương lai.

Chương 1

KIẾN THỨC CƠ SỞ

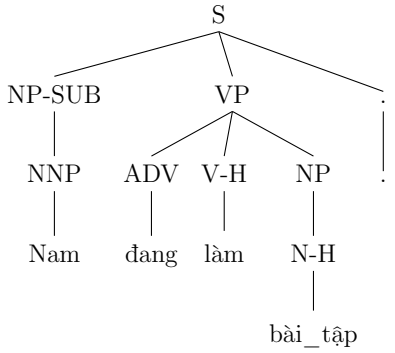
Chương này trình bày về các khái niệm và kiến thức cơ bản về cú pháp và ngữ nghĩa, các mô hình ngôn ngữ, quy trình xây dựng kho ngữ liệu và một số tài nguyên quan trọng trong việc phân tích và biểu diễn ngữ nghĩa.

1.1 Một số vấn đề cơ bản về cú pháp và ngữ nghĩa

1.1.1 Cú pháp

a) Cú pháp thành phần

Định nghĩa: Cú pháp thành phần là cấu trúc các thành phần câu theo thứ bậc, biểu diễn trật tự, cách thức ghép nối các từ, cụm từ của câu. Ví dụ, với một câu tiếng Việt: ‘‘`Nam đang làm bài_tập.`’’ sẽ được phân tích cú pháp thành phần như trong Hình 1.1.



Hình 1.1: *Cây cú pháp thành phần của câu: Nam đang làm bài_tập .*

b) Cú pháp phụ thuộc

Định nghĩa: Cú pháp phụ thuộc là cấu trúc cú pháp chứa các mục từ vựng nối với nhau bởi các quan hệ nhị phân không đối xứng gọi là sự phụ thuộc. Ví dụ, một số quan hệ cú pháp phụ thuộc của câu `"Nam đang làm bài_tập."` như: nsubj(Nam, làm), dobj(làm, bài_tập), ...

1.1.2 Ngữ nghĩa

a) Các thông tin ngữ nghĩa

Các thông tin cơ bản của một mô hình biểu diễn ngữ nghĩa được thể hiện qua: Sự kiện, các tham tố cốt lõi và không phải cốt lõi, vai nghĩa [13] và PropBank [67], đồng sở chỉ và tu từ, quan hệ thời gian (TimeML [51]), quan hệ không gian, quan hệ diễn ngôn, cấu trúc logic và suy luận. Mỗi thành phần này đều đóng vai trò quan trọng trong việc hiểu và xử lý ngôn ngữ.

b) Các mô hình và ngôn ngữ biểu diễn ngữ nghĩa

Một số loại ngôn ngữ biểu diễn ngữ nghĩa sẽ được định nghĩa theo các dạng: dựa vào logic (*Logic-based formalisms*) sử dụng mệnh đề logic bậc nhất (*First-order logic - FOL*), dựa vào đồ thị (*Graph-based formalims*), và ngôn ngữ lập trình (*Programm Languages - PLs*) [82].

1.2 Các phương pháp phân tích cú pháp và ngữ nghĩa

1.2.1 Phát biểu bài toán

Một bài toán phân tích cú pháp, ngữ nghĩa giới hạn trong câu có thể được phát biểu hình thức như sau: Câu đầu vào là một chuỗi n từ: $x = w_1, w_2, \ldots, w_n$. Thông thường, câu x sẽ được trải qua một số bước tiền xử lý như tách từ và gán nhãn từ loại. Đầu ra: Thông tin cú pháp, ngữ nghĩa của câu x theo mô hình hoặc định dạng cụ thể.

1.2.2 Các phương pháp phân tích cú pháp - ngữ nghĩa

Các phương pháp phân tích cú pháp - ngữ nghĩa thường được phân thành hai loại chính: Các phương pháp truyền thống (dựa vào luật, dựa vào thống kê, các phương pháp kết hợp) và các phương pháp dựa vào sử dụng mạng nơron [55].

1.2.3 Mô hình ngôn ngữ và biểu diễn văn bản

Phần này trình bày về một số mô hình ngôn ngữ cơ bản, được huấn luyện trước phổ biến như: n-grams, Word2vec, FastText, GloVe, BERT và các mô hình ngôn ngữ lớn như Llama, Gemini, GPT.

1.3 Một số vấn đề cơ bản về xây dựng ngữ liệu

1.3.1 Phương pháp luận

Phương pháp để xây dựng kho ngữ liệu thường theo một số bước cụ thể như: Xác định mục tiêu và phạm vi, thu thập văn bản có sẵn, ghi âm và chép lời, phân loại ngữ liệu, tiền xử lý dữ liệu, xây dựng cấu trúc và gán nhãn dữ liệu, kiểm duyệt và chuẩn hóa, cập nhật, bảo trì, chia sẻ và công bố dữ liệu.

1.3.2 Chuẩn hoá biểu diễn tài nguyên ngôn ngữ

Phần này trình bày về ISO.TC 37/SC 4[1] phát triển các tiêu chuẩn quốc tế cho việc quản lý các nguồn tài liệu ngôn ngữ, với mục đích cung cấp tiêu chuẩn cho việc chú giải và biểu diễn dữ liệu ngôn ngữ cơ bản.

[1] `https://www.iso.org/committee/297592.html`

1.4 Các tài nguyên ngôn ngữ

Phần này sẽ trình bày chi tiết về các tài nguyên từ vựng và các kho ngữ liệu đã được xây dựng.

1.4.1 Tài nguyên từ vựng

a) WordNet

WordNet [86] [41] là một cơ sở dữ liệu từ vựng tiếng Anh được nhóm thành tập hợp các lớp đồng nghĩa về nhận thức (*synsets*), mỗi lớp thể hiện một khái niệm riêng biệt. WordNet được phát triển cho các ngôn ngữ khác như tiếng Pháp [103], tiếng Trung Quốc [119] và tiếng Việt [96].

b) VerbNet

VerbNet [68] là mạng động từ, trong đó các động từ được xếp thành các lớp khác nhau dựa vào thuộc tính ngữ pháp và ngữ nghĩa của chính các động từ đó. VerbNet gồm hơn 5,800 động từ, được chia thành 270 nhóm, theo cách phân loại động từ của Beth Levin [72].

c) FrameNet

FrameNet [13] là một dự án xây dựng một cơ sở dữ liệu từ vựng, gồm hơn 200,000 câu được chú giải thủ công, được liên kết với hơn 1,200 khung ngữ nghĩa.

d) VCL

Tài nguyên từ vựng tiếng Việt lớn nhất là Từ điển tiếng Việt dùng cho máy tính (*Vietnamese Computational Lexicon* – VCL) [94]. Hiện tại, VCL chứa gần 42,000 mục từ, biểu diễn các thông tin: hình thái học, cú pháp học và ngữ nghĩa học.

1.4.2 Các kho văn bản có chú giải ngữ pháp, ngữ nghĩa

a) Kho ngữ liệu cú pháp thành phần

Kho ngữ liệu cú pháp thành phần gồm có 3 phần chính: gán nhãn từ loại, phân tích cú pháp thành phần và chú giải phát âm. Một số treebank như: Penn Treebank [88], ChineseTreebank [124], FrenchTreebank [9], VietTreebank [101].

b) Kho ngữ liệu phân tích cú pháp phụ thuộc đa ngôn ngữ

Các kho ngữ liệu được gán nhãn quan hệ phụ thuộc đa ngôn ngữ (*Universal Dependency - UD*[2]) được xây dựng bởi hơn 150 nhóm nghiên cứu với hơn 200 treebank cho các ngôn ngữ khác nhau.

[2]https://universaldependencies.org/

c) Kho ngữ liệu có gán nhãn vai nghĩa

PropBank [67] là kho ngữ liệu mở rộng từ Penn Treebank bằng việc chú giải vai nghĩa cho các động từ gồm ID ngữ cảnh và tham tố của nó được gán nhãn vai nghĩa. PropBank chú giải ngữ nghĩa cho khoảng 40,000 câu trong tập dữ liệu Penn Treebank.

d) Kho ngữ liệu gán nhãn ngữ nghĩa trừu tượng AMR

Mô hình biểu diễn ngữ nghĩa trừu tượng (AMR) [14], nắm bắt thông tin mô tả "ai làm gì cho ai" trong câu. Các ngôn ngữ khác như tiếng Tây Ban Nha [120], tiếng Hàn [24], tiếng Trung [73], ... cũng đã xây dựng các kho AMR với số lượng từ 1,000 đến 5,000 câu.

e) Kho ngữ liệu ngữ nghĩa UCCA

Mô hình biểu diễn ngữ nghĩa UCCA [7] chú giải sự khác biệt ngữ nghĩa và hướng tới mục đích trừu tượng hóa cấu trúc cú pháp cụ thể. Kho dữ liệu có chú giải UCCA gồm có 56,980 tokens, trong 148 đoạn văn từ các bài báo từ Wikipedia tiếng Anh.

f) Kho ngữ liệu ngữ nghĩa dựa vào cú pháp phụ thuộc

Mô hình biểu diễn ngữ nghĩa dựa vào cú pháp phụ thuộc DCS [99] xây dựng một hệ thống trả lời các câu hỏi từ ngôn ngữ tự nhiên bằng cách trình bày ngữ nghĩa của nó dưới dạng một hình thức logic và tính toán các câu trả lời từ một cơ sở dữ liệu có cấu trúc của các sự kiện.

g) Kho ngữ liệu ngữ nghĩa Groningen

Kho dữ liệu ngữ nghĩa Groningen (GMB) [58] bao gồm các văn bản tiếng Anh như các bài báo, tạp chí, ... với các biểu diễn cú pháp và biểu diễn ngữ nghĩa tương ứng. GMB được phát triển bởi nhóm nghiên cứu của trường Đại học Groningen, có phiên bản đa ngôn ngữ. Phiên bản cuối cùng gồm 10,000 văn bản với hơn 1 triệu từ loại.

1.5 Kết luận chương 1

Trong chương này, tác giả đã trình bày các kiến thức cơ sở cho luận án, cụ thể là:

- Những vấn đề cơ bản về phân tích cú pháp và ngữ nghĩa: cú pháp thành phần, cú pháp phụ thuộc, biểu diễn và phân tích ngữ nghĩa.
- Các phương pháp phân tích cú pháp và ngữ nghĩa.
- Các tiêu chuẩn xây dựng và chuẩn hoá biểu diễn tài nguyên ngôn ngữ.
- Các tài nguyên ngôn ngữ

Chương 2

XÂY DỰNG TÀI NGUYÊN VÀ CÔNG CỤ CHÚ GIẢI NGỮ PHÁP TIẾNG VIỆT

Chương này sẽ trình bày chi tiết các bước xây dựng các kho ngữ liệu cú pháp phụ thuộc, cú pháp thành phần và thuật toán chuyển đổi giữa hai kho ngữ liệu này cho tiếng Việt.

2.1 Kho ngữ liệu phân tích cú pháp phụ thuộc cho tiếng Việt (*viDependencyTreebank*)

2.1.1 Xây dựng tập nhãn cú pháp phụ thuộc tiếng Việt

Luận án thực hiện rà soát và xây dựng toàn bộ bộ nhãn phụ thuộc, xây dựng một bộ gồm 84 nhãn (40 nhãn chính và 44 nhãn con) dành cho tiếng Việt, dựa vào hướng dẫn gán nhãn của Phụ thuộc đa ngôn ngữ[1] phiên bản 2.11. Cụ thể các nhãn được mô tả trong Tài liệu hướng dẫn gán nhãn[2].

2.1.2 Kho ngữ liệu cú pháp phụ thuộc tiếng Việt

Kho ngữ liệu cú pháp phụ thuộc tiếng Việt đã xây dựng được mô tả trong Bảng 2.1.

Bảng 2.1: *Một số thống kê trên bộ dữ liệu cú pháp phụ thuộc tiếng Việt.*

Dữ liệu	Số câu	Độ dài <30	Độ dài 30-50	Độ dài >50	Độ dài Trung bình
Bộ huấn luyện Package1	5,069	4,882	159	28	14.40
Bộ huấn luyện Package2	3,083	1,942	1,005	136	24.96
Dữ liệu kiểm thử VLSP 2020	1,123	852	229	42	23.29
Dữ liệu kiểm thử mới	573	573	0	0	7.1

2.1.3 Thử nghiệm một số thuật toán phân tích cú pháp phụ thuộc

a) Xây dựng mô hình phân tích cú pháp phụ thuộc tiếng Việt

Luận án đã phát triển 8 mô hình để phân tích cú pháp phụ thuộc tiếng Việt, 6 mô hình trong số đó được xây dựng dựa vào mô hình deep bi-affine [38] và hai mô hình còn lại dựa trên mô hình con trỏ ngăn xếp (*Stack pointer*) [85]. Trong thực nghiệm, chúng tôi sử dụng nhiều biểu diễn phân bố từ khác nhau như Word2vec

[1] https://universaldependences.org/u/dep/index.html

[2] https://drive.google.com/file/d/1yEav7Nt4aw6wZvCiYb6rMxOZV9hiPoy-/view

[118], PhoBERT-base và PhoBERT-large [31], BARTPho [93] và XLM-RoBERTa [27], bên cạnh các phương pháp huấn luyện đa dạng kết hợp nhãn POS hoặc loại trừ chúng.

b) Độ đo đánh giá

Các mô hình phân tích cú pháp phụ thuộc được đánh giá bằng độ đo LAS (*Labeled Attachment Score*) và UAS (*Unlabeled Attachment Score*).

c) Đánh giá kết quả

Sau khi thử nghiệm các mô hình đã xây dựng, kết quả thu được là mô hình Deep-biaffine sử dụng PhoBERT đạt được độ chính xác cao nhất là 85.05%. Ngoài việc thử nghiệm với dữ liệu đã tách từ và gán nhãn từ loại, luận án còn thử nghiệm một số kịch bản khác như: dữ liệu đầu vào là dữ liệu thô, huấn luyện chỉ trên các nhãn phụ thuộc chính để so sánh các kết quả đạt được.

d) Thảo luận

Các yếu tố độ dài

Trong tiếng Việt, câu ngắn có độ chính xác cao hơn câu dài khoảng 2%.

Các yếu tố đồ thị

Luận án tập trung vào việc thống kê các lỗi liên quan tới khoảng cách đến gốc. Hầu hết các độ đo đều cao nhất với khoảng cách là 2 và 3.

Các yếu tố ngôn ngữ

Đối với các loại phụ thuộc, có thể thấy một số kiểu phụ thuộc phổ biến và không bị nhầm lẫn với các trường hợp khác như *root*, *obj*, *nsubj*, *case*, *cc*, *conj*, *advmod* sẽ có độ chính xác cao.

2.2 Kho ngữ liệu cú pháp thành phần cho tiếng Việt

2.2.1 Xây dựng tập nhãn cú pháp thành phần tiếng Việt

Việc xây dựng tập nhãn Viettreebank sẽ được thực hiện trên các phần: nhãn từ loại, nhãn cụm từ, nhãn chức năng cú pháp và nhãn mệnh đề.

a) Tách từ

Đối với việc tách từ, ngoài những tiêu chuẩn và khái niệm cơ bản, có một số thay đổi quan trọng được mô tả như: Đối với tên riêng, cụm từ (MWE).

b) Gán nhãn từ loại

Một số thay đổi trong nhãn từ loại như: nhãn X, nhãn Z, dấu câu. Một số nhãn từ loại mới cũng được thêm vào để có thể nắm bắt những đặc trưng của tiếng Việt: V:cop, V:mod, V:pass, ADJ:adv, ...

c) Nhãn ngữ đoạn

Đối với nhãn cụm, luận án đã xây dựng và thêm vào một số nhãn mới, điều này giúp cho việc phân biệt các cụm từ rõ ràng hơn.

d) Nhãn chức năng và nhãn mệnh đề

Tương tự như các nhãn trên, luận án cũng đã so sánh và xây dựng một số nhãn chức năng và mệnh đề mới dựa vào tập nhãn tiếng Anh.

2.2.2 Kho ngữ liệu cú pháp thành phần tiếng Việt

Các văn bản sử dụng trong VCP 2023 được thu thập từ 4 nguồn chính: kho ngữ liệu VTB, tập dữ liệu NER-VLSP 2021 [75], tập dữ liệu hồ sơ bệnh án điện tử (EMR) và một tập nhỏ các câu từ tin tức y tế trực tuyến. Bảng 2.1 thống kê các thông số trong tập dữ liệu VCP 2023 đã được xây dựng.

Bảng 2.2: *Thống kê dữ liệu VCP 2023.*

Nhãn	DL huấn luyện	DL kiểm thử 1	DL kiểm thử 2
Số câu	8,242	500	1,020
Độ dài trung bình	21	19	20
VP	31,800	1,933	4,017
NP	49,437	3,048	5,735
AP	5,980	440	974
PP	10,054	624	1,434
RP	948	27	180
WHADVP	56	3	16

2.2.3 Khảo sát các công cụ phân tích cú pháp thành phần cho tiếng Việt

Kết quả của các mô hình phân tích cú pháp thành phần được mô tả trong Bảng 2.3.

2.3 Thuật toán chuyển từ phân tích cú pháp thành phần sang cú pháp phụ thuộc

Luận án thực hiện xây dựng các luật để chuyển từ cú pháp thành phần sang cú pháp phụ thuộc và ngược lại.

2.3.1 Từ cú pháp thành phần sang cú pháp phụ thuộc

a) Xây dựng luật xác định từ trung tâm

Luận án đã xây dựng một bộ luật để tìm ra trung tâm của cụm từ dựa vào nghiên cứu [91]. Các luật xác định từ trung tâm (*headrules*) đã được cập nhật (sửa luật, thay thế nhãn) để phù hợp với tập nhãn trong kho ngữ liệu tiếng Việt hiện tại. Tập luật xác định từ trung tâm gồm có 21 luật.

Bảng 2.3: *Kết quả của các mô hình phân tích cú pháp thành phần.*

STT	Mô hình	DL huấn luyện	Word Embedding	Tối ưu hoá	F_1
1	Mô hình CRF hai giai đoạn	8,160	xlm-roberta-large, PhoBERT large	AdamW	83.46%
2	Mô hình mạng nơ-ron dựa vào phân tích cú pháp và tách từ Stanza	8,160	PhoBERT large	AdaDelta, Madgrad	83.93%
3	Mô hình mạng nơ-ron sử dụng Attach-juxtapose	8,242	Word2vec, PhoBERT base, large	AdamW	86.15%
4	Mô hình HPSG kết hợp với Stanza-tagger	8,242	PhoBERT large	AdaDelta, Madgrad	90.15%

b) Xây dựng luật xác định nhãn phụ thuộc

Sau khi đã xác định được các từ trung tâm, có nghĩa là đã xác định được hai từ trong một câu có mối quan hệ, thì việc tiếp theo chính là đặt tên cho mối quan hệ đó. Luận án đã xây dựng được bộ luật gồm khoảng 60 luật để xác định nhãn phụ thuộc của một mối quan hệ trong câu.

c) Kết quả

Kết quả của công cụ chuyển đổi từ CPTP sang CPPT được thực hiện trên 5,908 câu được mô tả trong Bảng 2.4.

Bảng 2.4: *Kết quả chuyển cú pháp thành phần sang cú pháp phụ thuộc.*

Số câu	LAS	UAS
5,908	52.63%	66.20%

2.3.2 Từ cú pháp phụ thuộc sang cú pháp thành phần

a) Xây dựng thuật toán

Luận án đã nghiên cứu và xây dựng được một bộ luật gồm 22 nhãn DP cần phải thêm nhãn cụm cho CP. Và khi chuyển đổi từ DP sang CP, có 14 nhãn DP cần được đánh dấu trọng tâm (-H) để đảm bảo tính chính xác và rõ ràng trong việc phân tích.

b) Kết quả

Luận án đã thực nghiệm công cụ chuyển từ CPPT sang CPTP với bộ dữ liệu gồm 8,152 câu tiếng Việt. Kết quả cụ thể được mô tả trong Bảng 2.5.

Bảng 2.5: *Kết quả chuyển cú pháp phụ thuộc sang cú pháp thành phần.*

Kịch bản	**Precision**	**Recall**	**F1-score**
Đánh giá thô	88.12%	74.66%	**80.83%**
Đánh giá chính xác	82.25%	71.04%	76.23%

2.4 Kết luận chương 2

Chương này trình bày quá trình xây dựng kho ngữ liệu cú pháp phụ thuộc và cú pháp thành phần cho tiếng Việt theo hướng tiếp cận đối sánh đa ngữ, phát triển và khảo sát các thuật toán phân tích cú pháp. Cuối cùng, luận án đề xuất thuật toán chuyển đổi giữa hai kho ngữ liệu và đánh giá kết quả đạt được.

Chương 3

XÂY DỰNG TÀI NGUYÊN VÀ CÔNG CỤ CHÚ GIẢI NGỮ NGHĨA TIẾNG VIỆT

Chương này sẽ trình bày về việc xây dựng tài nguyên và công cụ chú giải ngữ nghĩa tiếng Việt.

3.1 Kho ngữ liệu có gán nhãn vai nghĩa cho tiếng Việt theo cách tiếp cận liên ngữ

Việc xây dựng được tập nhãn vai nghĩa cho tiếng Việt được thực hiện dựa vào nhãn PropBank tiếng Anh, quy trình xây dựng kho ngữ liệu gán nhãn vai nghĩa [10] và các phiên bản PropBank của tiếng Việt trước đó. Bộ nhãn được xây dựng bao gồm 42 nhãn, được mô tả chi tiết trong Tài liệu gán nhãn PropBank[1].

Ngoài ra, việc xây dựng kho ngữ liệu có gán nhãn vai nghĩa còn được thực hiện dựa vào việc dóng hàng các khung PropBank từ tiếng Việt sang tiếng Anh. Kho ngữ liệu có gán nhãn vai nghĩa cho tiếng Việt đã được xây dựng gồm có 2,570 câu, từ hai tập dữ liệu Hoàng tử bé và Viettreebank, mô tả chi tiết trong Bảng 3.1.

Bảng 3.1: *Thống kê trên từng tập dữ liệu trong PropBank.*

Nhãn	**Hoàng Tử Bé**	**Viettreebank**
Số câu	1,570	1,000
Số từ	18,096	13,968
Vị từ (động từ)	2,278	2,018
Số lượng nhãn vai nghĩa	15,537	14,654
Tập nhãn	42	30

3.2 Mô hình biểu diễn và phân tích ngữ nghĩa cho tiếng Việt

3.2.1 Các mô hình vai nghĩa và mô hình biểu diễn ngữ nghĩa

Nhiều mô hình gán nhãn vai nghĩa đã được các nhóm nghiên cứu phát triển nhằm phục vụ cho các mục tiêu biểu diễn và phân tích ngữ nghĩa, điển hình là các mô hình như FrameNet, PropBank, và VerbNet. Mỗi mô hình có một cách

[1] https://docs.google.com/document/d/1g9PEDe2qgQ7jnTMkKzOX8dWUQURP87fP/edit

tiếp cận và mức độ chi tiết riêng biệt trong việc chú giải vai nghĩa, phản ánh các khía cạnh ngữ nghĩa của ngôn ngữ ở các cấp độ khác nhau. Luận án lựa chọn mô hình vai nghĩa LIRICS [98] và AMR [14] là cơ sở để xây dựng mô hình biểu diễn ngữ nghĩa cho tiếng Việt.

3.2.2 Xây dựng tập nhãn ngữ nghĩa tiếng Việt

Luận án đã lựa chọn AMR[2] là cơ sở để xây dựng mô hình biểu diễn ngữ nghĩa cho tiếng Việt. Luận án đã xây dựng mô hình gồm 18 vai chính, 71 vai phụ, 17 nhãn về thời gian, địa điểm và 4 nhãn về câu. Hướng dẫn gán nhãn chi tiết trong tệp: Hướng dẫn gán nhãn AMR tiếng Việt[3].

3.2.3 Xây dựng công cụ gán nhãn ngữ nghĩa cho tiếng Việt

Để xây dựng kho ngữ liệu biểu diễn ngữ nghĩa trừu tượng AMR cho tiếng Việt, luận án đã xây dựng ứng dụng web [4] để có thể gán nhãn dữ liệu một cách nhanh chóng và chính xác.

3.2.4 Kho ngữ liệu gán nhãn ngữ nghĩa cho tiếng Việt

Hiện tại, luận án đã xây dựng được kho ngữ liệu gán nhãn ngữ nghĩa cho tiếng Việt gồm 1,570 câu từ tiểu thuyết Hoàng Tử Bé. Kho ngữ liệu được gán nhãn bán thủ công trên công cụ chuyển đổi (dựa vào luật) và công cụ gán nhãn đã xây dựng trước đó.

Một số thống kê về tập nhãn của kho ngữ liệu được mô tả trong Bảng 3.2.

Bảng 3.2: *Thống kê 20 nhãn xuất hiện nhiều nhất trong kho dữ liệu ngữ nghĩa tiếng Việt.*

Nhãn	**Số lần xuất hiện**	**Nhãn**	**Số lần xuất hiện**
mod	1,376	polarity	341
agent	1,193	domain	338
theme	655	op2	330
compound	522	manner	309
quant	481	op1	296
classifier	433	patient	289
pivot	415	time	276
degree	412	and	247
topic	383	poss	236
polarity	341	tense	177

[2] https://github.com/amrisi/amr-guidelines/blob/master/amr.md

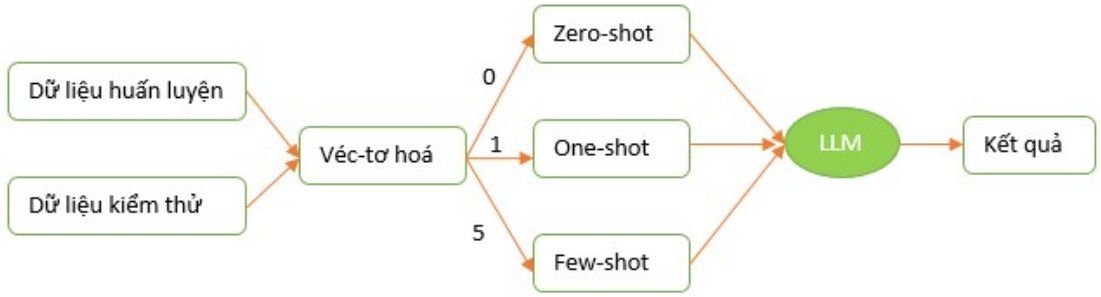
[3] https://docs.google.com/document/d/14t6DAZjwEkhXoJHFY6GVVpScFrZLgdzf/edit?usp=sharing&ouid=117154363694742364830&rtpof=true&sd=true

[4] https://amr.hpda.vn/login

3.3 Xây dựng mô hình phân tích ngữ nghĩa cho tiếng Việt

Với bài toán sinh biểu diễn ngữ nghĩa cho văn bản tiếng Việt, luận án đã thực hiện viết các prompt cho 2 mô hình ngôn ngữ lớn là Gemini và GPT-4, đồng thời để các mô hình học theo cả ba cách zero-shot, one-shot và few-shot.

Luồng công việc sử dụng mô hình ngôn ngữ lớn sinh biểu diễn ngữ nghĩa cho tiếng Việt được mô tả cụ thể trong Hình 3.1.



Hình 3.1: *Mô hình ngôn ngữ lớn sinh biểu diễn ngữ nghĩa cho tiếng Việt.*

3.3.1 Các độ đo đánh giá

Hệ thống gán nhãn vai nghĩa và phân tích ngữ nghĩa được đánh giá dựa trên độ chính xác (P), độ bao phủ (R), và $F_1 - score$ (với P và R được định nghĩa riêng cho từng bài toán).

3.3.2 Kết quả

a) Mô hình gán nhãn vai nghĩa

Mô hình GPT-4 được áp dụng để gán nhãn vai nghĩa cho tiếng Việt, sử dụng phương pháp *few-shot*. Kết quả được thể hiện trong Bảng 3.3.

Bảng 3.3: *Kết quả đánh giá mô hình ngôn ngữ lớn gán nhãn vai nghĩa cho tiếng Việt.*

Độ đo	Loại	Precision	Recall	F_1
conll09-head	Predicate	78.18%	78.18%	78.18%
conll09-head	ArgumentHead	53.99%	41.67%	47.04%
conll05-span	ArgumentSpan	50.13%	28.36%	36.22%

b) Mô hình phân tích ngữ nghĩa

Ngoài việc xây dựng prompt cho các mô hình ngôn ngữ lớn, luận án còn thử nghiệm mô hình Ensemble [71] đã xây dựng cho tiếng Anh với dữ liệu tiếng Việt,

gọi tên là *ViBART*. Luận án đã huấn luyện lại mô hình này trên kho dữ liệu biểu diễn ngữ nghĩa tiếng Việt và sử dụng biểu diễn phân bố từ BARTPho [93]. Kết quả chi tiết được mô tả trong Bảng 3.4.

Bảng 3.4: *Kết quả sinh biểu diễn ngữ nghĩa tiếng Việt.*

STT	Mô hình	Prompt	Smatch (F_1)	Số câu lỗi
1	ViBART	-	55.90%	0
2	GPT-4	Zero-shot	10%	16
		One-shot	47.88%	0
		5-shot	55.36%	0
		Few-shot	53.25%	0
3	Gemini	Zero-shot	16%	12
		One-shot	46.44%	1
		5-shot	57.72%	0
		Few-shot	54.90%	1

3.4 Kết luận chương 3

Chương này đã trình bày quá trình xây dựng kho ngữ liệu gán nhãn vai nghĩa và mô hình chú giải ngữ nghĩa cho tiếng Việt theo hướng tiếp cận liên ngữ. Kho ngữ liệu đã xây dựng được bao gồm 2,570 câu có gán nhãn vai nghĩa và 1,570 câu có gán nhãn ngữ nghĩa. Bên cạnh đó, luận án đã phát triển một công cụ hỗ trợ gán nhãn ngữ nghĩa và thử nghiệm xây dựng mô hình phân tích ngữ nghĩa cho tiếng Việt, đồng thời so sánh kết quả với các mô hình ngôn ngữ lớn hiện có. Mặc dù kết quả đạt được còn khiêm tốn, nhưng đây là mô hình biểu diễn ngữ nghĩa sâu đầu tiên dành riêng cho tiếng Việt. Kho ngữ liệu gán nhãn và các mô hình được phát triển trong luận án sẽ là nền tảng quan trọng cho các nghiên cứu sâu hơn về ngữ nghĩa trong tiếng Việt.

Chương 4

XÂY DỰNG MẠNG ĐỘNG TỪ TIẾNG VIỆT

Ở chương này, luận án sẽ trình bày những nghiên cứu và các công việc cụ thể để xây dựng VerbNet cho tiếng Việt (*viVerbNet*).

4.1 Từ điển tiếng Việt cho máy tính VCL

Đối với tiếng Việt, từ điển tiếng Việt cho máy tính VCL [94] là một tài nguyên từ vựng duy nhất và hữu ích trong nghiên cứu ngữ pháp, ngữ nghĩa. Luận án thực hiện khảo sát và so sánh VCL với VerbNet theo các tiêu chuẩn: về cách thức tổ chức, về các thông tin biểu diễn (thông tin hình thái, cú pháp, khung vị từ, vai nghĩa, ràng buộc lựa chọn, ràng buộc cú pháp và vị từ ngữ nghĩa). Những thông tin biểu diễn trong VCL được sử dụng để xây dựng viVerbNet.

4.2 Phương pháp xây dựng viVerbNet

Luận án đã thực hiện phân cụm động từ tiếng Việt thành các nhóm động từ. Động từ sẽ được trích xuất từ VCL, sau đó tìm kiếm trong kho ngữ liệu để lấy ngữ cảnh. Khi đã có ngữ cảnh của các động từ, luận án sử dụng một số mô hình biểu diễn véctơ từ (*word embedding*) đã được huấn luyện bằng các ngữ liệu tiếng Việt để sinh véc tơ từ cho các động từ này, đây sẽ là đầu vào cho các thuật toán phân cụm. Sau khi phân cụm động từ, luận án sử dụng các thông tin ngữ pháp tiếng Việt, kết hợp với trích xuất một số thông tin từ các kho ngữ liệu chú giải cú pháp và ngữ nghĩa tiếng Việt, đồng thời ánh xạ VerbNet tiếng Anh để xây dựng các thành phần của một cụm động từ: khung cú pháp, vai nghĩa, ràng buộc lựa chọn, ràng buộc cú pháp và vị từ ngữ nghĩa.

4.2.1 Biểu diễn véc-tơ từ

Luận án sử dụng véc tơ biểu diễn từ làm đầu vào cho bài toán phân cụm động từ tiếng Việt: Mô hình Word2vec[87] [118], Mô hình PhoBERT [31].

4.2.2 Phân cụm động từ tiếng Việt

Đối với bước phân cụm động từ, hai thuật toán được sử dụng là K-means và HCA. Luận án chọn 1,000 làm số lượng các cụm.

4.3 Xây dựng các thành phần của viVerbNet

4.3.1 Vai nghĩa

Luận án sử dụng 24 vai nghĩa từ Propbank tiếng Việt và 29 vai nghĩa LIRICS như một cơ sở để phát triển các vai nghĩa cho viVerbNet.

4.3.2 Ràng buộc lựa chọn

Luận án sử dụng những ràng buộc lựa chọn để chỉ ra sự tồn tại (+) hoặc không tồn tại (-) của các thuộc tính ngữ nghĩa như [concrete], [animate], [organization], ... Các toán tử logic (| (hoặc) và & (và)) được sử dụng để kết hợp nhiều hạn chế.

4.3.3 Khung cú pháp và ràng buộc cú pháp

Khung cú pháp trong VerbNet mô tả ngắn gọn cấu trúc bề mặt của các thành phần cấu thành câu. Nó bao gồm các vai nghĩa tương ứng với các tham thể, động từ chính và các ràng buộc về cú pháp. Luận án sẽ đối chiếu điểm khác biệt đó đồng thời đưa ra các lý giải, giải pháp để xây dựng viVerbNet.

4.3.4 Vị từ ngữ nghĩa

Các vị từ ngữ nghĩa biểu thị mối quan hệ giữa tham thể và các sự kiện để biểu thị ý nghĩa chính của câu. Thông tin ngữ nghĩa cho các động từ trong VerbNet được thể hiện dưới dạng kết hợp của các các vị từ ngữ nghĩa, chẳng hạn như vị từ ngữ nghĩa chung (chuyển động (*motion*), liên hệ (*contact*), truyền đạt thông tin (*transfer_info*), ...), vị ngữ (*Prep, Adv, và Pred*), vị ngữ cụ thể; vị ngữ cho nhiều sự kiện.

Đối với thành phần ngữ nghĩa trong viVerbNet, luận án sử dụng cùng một tập hợp các vị từ ngữ nghĩa như VerbNet. Tập hợp vị từ ngữ nghĩa của VerbNet tiếng Anh được luận án đồng sử dụng trong viVerbNet có khoảng 153 nhãn vị từ ngữ nghĩa.

4.4 Công cụ gán nhãn mạng động từ tiếng Việt

Để việc gán nhãn các lớp động từ trong tiếng Việt được đơn giản và đỡ tốn thời gian, công sức, luận án đã thiết kế công cụ gán nhãn mạng động từ tiếng Việt. Công cụ này được xây dựng để có thể sử dụng các kết quả từ các nghiên cứu trước đó, giúp việc gán nhãn các lớp động từ tiếng Việt nhanh gọn và chính xác hơn.

4.5 Ví dụ một cụm động từ trong viVerbNet

Để làm rõ việc xây dựng viVerbNet, luận án sẽ mô tả kĩ về một cụm động từ đã xây dựng đầy đủ các thành phần. Cụm động từ được lựa chọn được mô tả trong Bảng 4.1.

Bảng 4.1: *Nhóm động từ "học" trong viVerbNet.*

STT	**Nhóm động từ**	**Lớp**	**Lớp con**	**Chi tiết**
1	Nghĩa "***học***"	học-1	học-1.1	học, học hành, học tập, luyện tập, luyện, ôn, ôn luyện, ôn tập, tập luyện
			học-1.2	bắt chước, học lỏm, nhái, nhại
			học-1.3	học hỏi, chắt lọc, định hình, lĩnh hội, lãnh hội, rèn giũa, mài giũa, thu nạp, tích lũy, tiếp thụ, trau dồi

4.5.1 Vai nghĩa

Các vai nghĩa trong lớp này gồm có: đối tượng học (Agent), nguồn truyền đạt (Source), nội dung (Topic).

```
Role: Agent [+animate], Topic, Source
```

4.5.2 Ràng buộc lựa chọn

Về ràng buộc lựa chọn cho vai nghĩa, lớp "learn-14" trong VerbNet sử dụng duy nhất một ràng buộc [+animate]/[+tính động].

```
Role: Agent [+human], Topic, Source
```

4.5.3 Khung cú pháp và ràng buộc cú pháp

Khung cú pháp và ràng buộc của lớp "learn-14" và các lớp con của nó được biểu diễn như sau:

- learn-14

  ```
  (1) NP V NP PP. source
  syntax Agent V Topic from Source
  ```

4.5.4 Vị từ ngữ nghĩa

Các vị từ được sử dụng: vị từ chung "Transfer_infor" và vị từ quá trình "During(E)". Các ngữ nghĩa của lớp "học-1" và các lớp con được biểu thị như sau:

- Agent V Topic from Source: Tôi học tiếng Anh từ anh trai
  `Transfer_info(During(E), Source, Agent, Topic)`
- Agent V from Source: Tôi học từ anh trai tôi

Chi tiết về các cụm động từ được mô tả trong Dữ liệu phân cụm động từ của Luận án[1].

4.6 Kết luận chương 4

Chương này đã trình bày chi tiết về quá trình xây dựng mạng động từ viVerbNet cho tiếng Việt, một hệ thống nhằm phân loại và nhóm các động từ theo ngữ nghĩa và cú pháp một cách có hệ thống. Trước tiên, việc khảo sát kho từ vựng VCL và VerbNet đã được thực hiện, tiếp theo là quá trình phân cụm các động từ trong VCL và xây dựng được 100 cụm động từ cơ bản cho tiếng Việt. Mỗi cụm động từ đều được phát triển đầy đủ với các thành phần quan trọng như vai nghĩa, khung cú pháp với các ràng buộc cú pháp và ngữ nghĩa, thông tin về vị từ ngữ nghĩa. Mặc dù kết quả này vẫn chưa bao quát hết toàn bộ các động từ tiếng Việt, nhưng đã đặt nền móng quan trọng cho việc xây dựng hệ thống mạng động từ trong tương lai, hỗ trợ cho việc phát triển các mô hình biểu diễn và phân tích ngữ nghĩa sâu hơn trong ngữ cảnh tiếng Việt.

[1] `https://drive.google.com/drive/folders/1LeJyKHHBuv_JwwN8V2njmXh9c5g4t8oG?usp=sharing`

KẾT LUẬN

Luận án tập trung nghiên cứu các phương pháp biểu diễn và phát triển ngữ liệu, công cụ cho bài toán phân tích cú pháp và ngữ nghĩa tiếng Việt. Cụ thể, luận án đã có những đóng góp cơ bản về hai hướng chính:

- Xây dựng các tài nguyên ngôn ngữ, gồm các lược đồ chú giải, hướng dẫn chú giải và kho ngữ liệu có chú giải theo lược đồ đã thiết kế:
  - Cú pháp phụ thuộc: Trên cơ sở tập nhãn cú pháp phụ thuộc tiếng Việt đã xây dựng trong giai đoạn trước, luận án tiến hành cập nhật, thiết kế lại và chỉnh sửa tập nhãn cũng như hướng dẫn chú giải theo phiên bản 2.0 của Dự án cú pháp phụ thuộc phổ quát (*Universal Dependency - UD*). Sau đó, luận án tiến hành thu thập và xây dựng kho ngữ liệu với hơn 9,000 câu (trong đó 3,000 câu đã được tích hợp vào kho UD hồi tháng 11 năm 2022) đã được chú giải theo quy trình chuẩn hóa, với độ đồng thuận gán nhãn đạt 91%, đảm bảo chất lượng và tính nhất quán. Kho ngữ liệu này cùng với tài liệu hướng dẫn chi tiết đã được công khai trên GitHub[2] và sử dụng trong cuộc thi về phân tích cú pháp phụ thuộc tiếng Việt tại hội thảo về Xử lí ngôn ngữ tự nhiên và tiếng nói tiếng Việt (VLSP 2020).
  - Cú pháp thành phần: Kế thừa kho ngữ liệu cú pháp thành phần Viettreebank, luận án thực hiện việc rà soát, cập nhật và chuẩn hoá các nhãn cú pháp thành phần cũng như tài liệu hướng dẫn chú giải để có một bộ nhãn phù hợp với các nghiên cứu đối sánh đa ngữ. Trên cơ sở đó, kho ngữ liệu gồm hơn 9,000 câu đã được cập nhật theo tập nhãn mới với độ đồng thuận của các nhà chú giải lên tới 94% cho thấy kho ngữ liệu được xây dựng tỉ mỉ, chính xác và có độ tin cậy cao. Kho ngữ liệu này đã được công khai và sử dụng trong cuộc thi về phân tích cú pháp thành phần tiếng Việt tại hội thảo về Xử lý ngôn ngữ tự nhiên và tiếng nói tiếng Việt (VLSP 2022 và VLSP 2023).
  - Ngữ nghĩa nông (Vai nghĩa): Tập nhãn vai nghĩa cho tiếng Việt xây dựng trước đó đã được cập nhật và chỉnh sửa tương thích với khung chú giải vai nghĩa trong dự án Universal Proposition Bank 2.0. Tập nhãn này đã được sử dụng để xây dựng kho ngữ liệu tiếng Việt có chú giải vai nghĩa gồm 2,570 câu.
  - Ngữ nghĩa sâu: Luận án đã thiết kế mô hình biểu diễn ngữ nghĩa cho tiếng Việt dựa trên AMR và tập vai nghĩa LIRICS. Mô hình nổi bật nhờ khả năng thể hiện đặc trưng tiếng Việt như danh từ hóa động từ, ngữ nghĩa thời gian và đồng sở chỉ liên đoạn. Kho ngữ liệu gồm 1.570 câu từ tiểu thuyết Hoàng

[2] `https://github.com/vietnamesedp/Thesis`

tử bé được xây dựng theo quy trình chuẩn hóa, với độ đồng thuận cao giữa các chuyên gia[3].

  - Mạng động từ tiếng Việt: Sau khi xây dựng các kho ngữ liệu và mô hình phân tích, luận án phát triển mạng động từ tiếng Việt (*viVerbNet*) dựa trên 100 cụm động từ tiêu biểu, với 5 thành phần chính: vai nghĩa, ràng buộc lựa chọn, khung và ràng buộc cú pháp, vị từ ngữ nghĩa. Các lớp động từ được ánh xạ sang VerbNet tiếng Anh, góp phần kết nối tài nguyên ngữ nghĩa song ngữ và hỗ trợ tích hợp tiếng Việt vào hệ thống đa ngữ.

- Về phương pháp và mô hình cho phân tích tiếng Việt, luận án đã thực hiện những công việc sau:
  - Đánh giá, so sánh các mô hình véc-tơ từ huấn luyện sẵn cho tiếng Việt và một số phương pháp hiện đại để cải tiến hiệu quả của bài toán phân tích cú pháp. Cụ thể, với cú pháp thành phần, kết quả tốt nhất đạt $F_1 = 90.15\%$ với mô hình HPSG kết hợp với công cụ gán nhãn Stanza. Đối với cú pháp phụ thuộc, mô hình Deep bi-affine sử dụng PhoBERT đạt $LAS = 78.05\%$ và $UAS = 85.27\%$ - kết quả tốt nhất khi huấn luyện và kiểm thử trên kho ngữ liệu cú pháp phụ thuộc đã được xây dựng.
  - Xây dựng công cụ chuyển đổi giữa cú pháp thành phần và cú pháp phụ thuộc, hỗ trợ quá trình gán nhãn dữ liệu: Thuật toán chuyển đổi từ cú pháp thành phần sang cú pháp phụ thuộc đạt kết quả $LAS = 52.63\%$ và $UAS = 66.20\%$. Với chiều ngược lại, thuật toán chuyển đổi đạt kết quả $F_1 = 80.83\%$. Công cụ này cho phép linh hoạt chuyển đổi giữa hai cách biểu diễn ngữ pháp, được sử dụng trong việc xây dựng kho ngữ liệu, giúp giảm thời gian và tiết kiệm công sức của các chuyên gia chú giải.
  - Phát triển và đánh giá các thuật toán phân cụm động từ tiếng Việt: Các thuật toán phân cụm được thực hiện trên hơn 12,000 nghĩa của động từ trích từ từ điển tiếng Việt cho máy tính VCL, nhằm nhóm các động từ thành các cụm có chung những đặc điểm về ngữ pháp, ngữ nghĩa. Thuật toán phân cụm K-means đã được thử nghiệm chi tiết với nhiều kịch bản khác nhau, bao gồm thay đổi số lượng cụm, các mô hình véc tơ từ và các câu ngữ cảnh của động từ.
  - Thử nghiệm áp dụng các mô hình ngôn ngữ lớn như GPT-4 và Gemini cho nhiệm vụ gán nhãn vai nghĩa và phân tích ngữ nghĩa tiếng Việt. Kết quả cho thấy các mô hình đạt độ chính xác từ 47% đến 58% trên cả hai bài toán, phản

[3] `https://github.com/vietnamesedp/Thesis/tree/main/MeaningRepresentation`

ánh tiềm năng ứng dụng của các mô hình này trong xử lý ngôn ngữ tự nhiên. Những kết quả này không chỉ minh chứng cho hiệu quả ban đầu của các mô hình ngôn ngữ lớn đối với tiếng Việt, mà còn giảm thiểu chi phí và công sức gán nhãn thủ công, đồng thời thúc đẩy quá trình phát triển các ứng dụng ngôn ngữ thông minh trong thực tiễn.

Trong tương lai, các hướng phát triển của luận án tập trung vào:

- Tiếp tục phát triển và mở rộng các kho ngữ liệu ngữ nghĩa, chia sẻ và công bố rộng rãi các tài nguyên này trong cộng đồng xử lý ngôn ngữ tiếng Việt: Luận án sẽ tiếp tục mở rộng quy mô và phạm vi của kho ngữ liệu gán nhãn ngữ nghĩa cho tiếng Việt, bổ sung thêm nhiều dạng ngữ nghĩa phức tạp và ngữ cảnh sử dụng.
- Phát triển tiếp mạng động từ viVerbNet và tăng cường khả năng liên thông liên ngữ: Luận án tiếp tục hoàn thiện viVerbNet bằng cách mở rộng lớp động từ, tinh chỉnh vai nghĩa, khung cú pháp và ràng buộc lựa chọn để nâng cao độ chính xác và tính ứng dụng. Việc ánh xạ với VerbNet tiếng Anh cũng được điều chỉnh nhằm xử lý khác biệt ngôn ngữ, tăng khả năng tích hợp tiếng Việt vào các hệ thống đa ngữ.
- Tinh chỉnh và khai thác các mô hình ngôn ngữ lớn cho các bài toán cú pháp và ngữ nghĩa tiếng Việt: Luận án sẽ tiếp tục nghiên cứu và cải tiến các phương pháp khai thác mô hình ngôn ngữ lớn (LLMs) như GPT-4, Gemini, ... cho các tác vụ phân tích cú pháp và ngữ nghĩa. Việc tinh chỉnh các mô hình này trên dữ liệu tiếng Việt và tích hợp với các tài nguyên ngôn ngữ đã được xây dựng sẽ giúp nâng cao chất lượng gán nhãn, giảm thiểu chi phí thủ công, đồng thời mở rộng khả năng ứng dụng của các mô hình trong các hệ thống thực tế như dịch máy, tóm tắt văn bản, hỏi đáp và trợ lý ảo.
- Cải thiện chất lượng kho ngữ liệu:
  - Đánh giá và điều chỉnh sơ đồ chú giải để tăng tính nhất quán và khả năng sử dụng lại. Xây dựng bộ tiêu chí đánh giá chất lượng chú giải và đề xuất các công cụ hỗ trợ hiệu chỉnh bán tự động.
  - Tiếp tục khảo sát và phân tích các hiện tượng ngôn ngữ đặc thù của tiếng Việt trong ngữ liệu, từ đó đề xuất các biểu diễn ngữ nghĩa phù hợp hơn cho các cấu trúc cú pháp–ngữ nghĩa độc đáo của tiếng Việt.

Tóm lại, luận án đã có những đóng góp quan trọng trong việc xây dựng một hệ thống tài nguyên ngôn ngữ tiếng Việt phong phú, đa dạng và được chú giải ở mức độ sâu, tập trung vào kho từ vựng động từ cùng các ngữ liệu chú giải cú pháp – ngữ nghĩa. Đồng thời, luận án cũng đã tiến hành nghiên cứu, phát triển và đánh giá các công cụ phân

tích cú pháp và ngữ nghĩa tiên tiến, được tối ưu hóa phù hợp với đặc thù của tiếng Việt. Việc công bố rộng rãi các tài nguyên và công cụ này không chỉ góp phần thúc đẩy sự phát triển bền vững và lâu dài của lĩnh vực xử lý ngôn ngữ tự nhiên tiếng Việt, mà còn tạo nền tảng vững chắc cho các hướng nghiên cứu tiếp theo trong tương lai.

TÀI LIỆU THAM KHẢO

Tài liệu tiếng Việt

[1] Lâm Nhựt Khang, Võ Lê Minh Trung, Nguyễn Huỳnh Hữu Đức (2017), *Xây dựng WordNet cho tiếng Việt*, FAIR, Đà Nẵng, Việt Nam, trang 1007-1014.

[2] Hà Mỹ Linh, Nguyễn Thị Lương, Nguyễn Việt Hùng, Nguyễn Thị Minh Huyền, Lê Hồng Phương, Phan Thị Huê (2014), *Xây dựng kho ngữ liệu mẫu có gán nhãn vai nghĩa cho tiếng Việt*, Hội thảo quốc gia lần thứ XVII: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông, Đắk-lăk, Việt Nam, trang 409-414.

[3] Nguyễn Lê Minh, Hoàng Thị Điệp, and Trần Mạnh Kế (2008), *Nghiên cứu luật hiệu chỉnh kết quả dùng phương pháp MST phân tích cú pháp phụ thuộc tiếng việt*, Kỷ yếu Hội thảo ICT.rda'08, trang 258-267.

[4] Nguyễn Minh Thuyết, Nguyễn Văn Hiệp (1998), *Thành phần câu tiếng Việt*, Nxb Đại học Quốc gia Hà Nội, Hà Nội.

[5] Nguyễn Thiện Giáp (2014), *Nghĩa học Việt ngữ*, Nhà xuất bản Giáo dục Việt Nam, 2014, 327 trang.

[6] Nguyễn Thị Bích Ngoan (2013), *So sánh đối chiếu hiện tượng danh hoá động từ trong tiếng Việt và tiếng Anh*, Tạp chí Khoa học Đại học Sư phạm TPHCM, trang 13–22.

Tài liệu tiếng Anh

[7] Abend, Omri and Rappoport, Ari (2013), *Universal Conceptual Cognitive Annotation (UCCA)*, Proceedings of ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 228–238.

[8] Abend Omri and Rappoport Ari (2017), *The State of the Art in Semantic Representation*, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 77–89.

[9] Abeillé, Anne and Clément, Lionel and Toussenel, François (2003), *Building a Treebank for French*, in Treebanks: Building and Using Parsed Corpora, edited by Anne Abeillé, Springer Netherlands, Dordrecht, pp. 165–187, ISBN: 978-94-010-0201-1, DOI: 10.1007/978-94-010-0201-1_10.

[10] Akbik, Alan and Chiticariu, Laura and Danilevsky, Marina and Li, Yunyao and Vaithyanathan, Shivakumar and Zhu, Huaiyu (2015), *Generating High Quality Proposition Banks for Multilingual Semantic Role Labeling*, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Vol. 1: Long Papers), pp 397–407.

[11] Artstein Ron and Poesio Massimo (2008), *Survey Article: Inter-Coder Agreement for Computational Linguistics*, Computational Linguistics, Vol. 34, No. 4, pp 555–596.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017), *Attention Is All You Need*, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[13] Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998), *The Berkeley FrameNet Project*, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Vol. 1, pp 86–90.

[14] Banarescu, Laura, Bonial, Claire, Cai, Shu, Georgescu, Madalina, Griffitt, Kira, Hermjakob, Ulf, Knight, Kevin, Koehn, Philipp, Palmer, Martha, and Schneider, Nathan (2013), *Abstract Meaning Representation for Sembanking*, Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp. 178–186.

[15] Bob Carpenter (1997), *Type-logical semantics*, MIT Press, Cambridge.

[16] Bojanowski, Piotr and Grave, Edouard and Joulin, Armand and Mikolov, Tomas (2017), *Enriching Word Vectors with Subword Information*, Transactions of the Association for Computational Linguistics, Vol. 5, pp. 135–146.

[17] Brown Peter F. and Della Pietra Vincent J. and deSouza Peter V. and Lai Jenifer C. and Mercer Robert L. (1992), *Class-Based n-gram Models of Natural Language*, Computational Linguistics, Vol 18, No. 4, pp 467–480.

[18] Brown, Susan and Dligach, Dmitriy and Palmer, Martha (2011), *VerbNet class assignment as a WSD task*, In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pp. 85–94.

[19] Cai, Shu and Knight, Kevin (2013), *Smatch: an Evaluation Metric for Semantic Feature Structures*, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria, Association for Computational Linguistics, pp. 748–752.

[20] Candito, Marie and Amsili, Pascal and Barque, Lucie and Benamara, Farah and de Chalendar, Gaël and Djemaa, Marianne and Haas, Pauline and Huyghe, Richard and Mathieu, Yvette Yannick and Muller, Philippe and Sagot, Benoît and Vieu, Laure (2014), *Developing a French FrameNet: Methodology and First results*, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp 1372–1379.

[21] Carl Pollard and Ivan A. Sag (1994), *Head-Driven Phrase Structure Grammar*, The University of Chicago Press, Chicago.

[22] Carreras Xavier and Màrquez Lluís (2004), *Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling*, in *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, Association for Computational Linguistics, pp 89–97.

[23] Carreras Xavier and Màrquez, Lluís (2005), *Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling*, in Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), Association for Computational Linguistics, pp 152–164.

[24] Choe, Hyonsu and Han, Jiyoon and Park, Hyejin and Oh, Tae Hwan and Kim, Hansaem (2020), *Building Korean Abstract Meaning Representation Corpus*, Proceedings of the Second International Workshop on Designing Meaning Representations, pp. 21–29.

[25] Clusmann Jan and Kolbinger Fiona and Muti Hannah and Carrero Zunamys and Eckardt Jan-Niklas and Laleh Narmin and Löffler Chiara and Schwarzkopf Sophie-Caroline and Unger Michaela and Veldhuizen, Gregory and Wagner, Sophia and Kather, Jakob (2023), *The future landscape of large language models in medicine*, Communications medicine, Vol. 3, No. 1, pp 1–8, doi: https://doi.org/10.1038/s43856-023-00370-1.

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020), *Exploring the Limits of*

Transfer Learning with a Unified Text-to-Text Transformer, Journal of Machine Learning Research, Vol. 21, No. 140, pp. 1–67.

[27] Conneau Alexis and Khandelwal Kartikay and Goyal Naman and Chaudhary Vishrav and Wenzek Guillaume and Guzman Francisco and Grave Edouard and Ott Myle and Zettlemoyer Luke and Stoyanov Veselin (2020), *Unsupervised Cross-lingual Representation Learning at Scale*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 8440–8451.

[28] Crystal, D. (1997), *A dictionary of linguistics and phonetics*, 4th edition, Cambridge, MA: Blackwell Publishing.

[29] Dat Quoc Nguyen, Nguyen Dai, Pham Son, Nguyen Phuong-Thai, and Nguyen Le (2014), *From Treebank Conversion to Automatic Dependency Parsing for Vietnamese*, Proceedings of the International Conference, pp. 196–207.

[30] Dat Quoc Nguyen, Mark Dras, and Mark Johnson (2016), *An empirical study for Vietnamese dependency parsing*, Proceedings of the 14th Annual Workshop of the Australasian Language Technology Association, ALTA 2016, pp 143–149.

[31] Dat Quoc Nguyen and Nguyen Anh Tuan (2020), *PhoBERT: Pre-trained language models for Vietnamese*, Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1037–1042.

[32] Day, William H. E. and Edelsbrunner, Herbert (1984), *Efficient algorithms for agglomerative hierarchical clustering methods*, *Journal of Classification*, Vol. 1, pp 7-24.

[33] Danlos Laurence and Nakamura Takuya and Pradet Quentin (2014), *Vers la création d'un VerbeNet du français*, Atelier FondamenTAL, TALN 2014, July, pp 103–108.

[34] Deng Jiawen and Zubair Areeba and Park Yejean (2023), *Limitations of large language models in medical applications*, Postgraduate Medical Journal, pp 1298-1299.

[35] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina (2018), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, NAACL-HLT 2019, pp 4171-4186.

[36] Diem Truong, Duc-Thuan Vo, and Uyen Trang Nguyen (2017), *Vietnamese Open Information Extraction*, In Proceedings of the 8th International Symposium on Information and Communication Technology (SoICT '17), Association for Computing Machinery, New York, NY, USA, pp 135–142, https://doi.org/10.1145/3155133.3155171.

[37] Dorr, Bonnie J. (1997), *Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation*, Machine Translation, Vol. 12, No. 4, pp 271–322.

[38] Dozat Timothy and Manning Christopher D. (2017), *Deep Biaffine Attention for Neural Dependency Parsing*, In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, Conference Track Proceedings, OpenReview.net.

[39] Emily M. Bender, Batya Friedman (2018), *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science*, Transactions of the Association for Computational Linguistics, Vol. 6, pp. 587–604.

[40] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen (2018), *CARER: Contextualized Affect Representations for Emotion Recognition*, In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3687–3697, Brussels, Belgium.

[41] Fellbaum, C. (Ed.). (1998), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA. ISBN: 978-0-262-06197-1.

[42] Francis, W. Nelson, and Kucera, Henry (1979), *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*, Brown University.

[43] Gemini Team and Rohan Anil and Sebastian Borgeaud and Jean-Baptiste Alayrac and Jiahui Yu and Radu Soricut, et al. (2024), *Gemini: A Family of Highly Capable Multimodal Models*, arXiv, url=https://arxiv.org/abs/2312.11805.

[44] Giuglea, Ana-Maria and Moschitti, Alessandro (2006), *Semantic Role Labeling via FrameNet, VerbNet and PropBank*, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 929–936.

[45] Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, Claudia Soria (2006), *Lexical Markup Framework (LMF)*, Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), pp. 577–580.

[46] Hajič Jan and Ciaramita, Massimiliano and Johansson, Richard and Kawahara, Daisuke and Martí, Maria Antònia and Màrquez, Lluís and Meyers, Adam and Nivre, Joakim and Padó, Sebastian and Štěpánek, Jan and Straňák, Pavel and Surdeanu, Mihai and Xue, Nianwen and Zhang, Yi (2009), *The CoNLL-2009 shared task: syntactic and semantic dependencies in multiple languages*, in Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task, Association for Computational Linguistics, pp 1–18.

[47] Hao Shibo and Gu Yi and Ma Haodi and Hong Joshua and Wang Zhen and Wang Daisy and Hu Zhiting (2023), *Reasoning with Language Model is Planning with World Model*, in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 8154–8173.

[48] Hartigan, J. A. and Wong, M. A. (1979), *Algorithm AS 136: A K-means clustering algorithm*, Applied Statistics, Royal Statistical Society, pp. 100–108.

[49] Hensman, Svetlana and Dunnion, John (2004), *Automatically building conceptual graphs using VerbNet and WordNet*, Proceedings of the Intenational Symposium on Information and Communication Technologies, Las Vegas, Nevada, USA, June 16-18, 2004, pp 115–120.

[50] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp 4171–4186.

———————————-

[51] James Pustejovsky and José M. Castaño and Robert Ingria and Roser Sauri and Robert J. Gaizauskas and Andrea Setzer and Graham Katz and Dragomir R. Radev (2003), *TimeML: Robust Specification of Event and Temporal Expressions in Text*,

In Mark T. Maybury (ed.), *New Directions in Question Answering, Papers from 2003 AAAI Spring Symposium*, Stanford University, Stanford, CA, USA, pp 28–34.

[52] Jay Earley (1970), *An efficient context-free parsing algorithm*, Communications of the ACM, 13(2), pp. 94–102. DOI: https://doi.org/10.1145/362007.362035

[53] Jiangming Liu and Yue Zhang (2016), *Shift-Reduce Constituent Parsing with Neural Lookahead Features*, Transactions of the Association for Computational Linguistics, Vol 5, pp 45–58, http://arxiv.org/abs/1612.00567

[54] Jiangming Liu and Yue Zhang (2017), *In-Order Transition-based Constituent Parsing*, *Transactions of the Association for Computational Linguistics*, Vol. 5, MIT Press, Cambridge, MA, pp. 413–424.

[55] Jiang Peng and Cai Xiaodong (2024), *A Survey of Semantic Parsing Techniques*, Symmetry, Vol. 16, No. 9, Article 1201.

[56] Jinqi Lai and Wensheng Gan and Jiayang Wu and Zhenlian Qi and Philip S. Yu (2024), *Large language models in law: A survey*, AI Open, Vol. 5, pp 181–196.

[57] Ji Tao, Wu Yuanbin, and Lan Man (2019), *Graph-based Dependency Parsing with Graph Neural Networks*, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2475–2485, Florence, Italy.

[58] Johan Bos (2013), *The Groningen Meaning Bank*, in Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora, edited by Octavian Popescu and Alberto Lavelli, Trento, Italy, Vol. 2, pp 463–496, https://aclanthology.org/W13-3802.

[59] Johnson, Tim (1984), *Natural language computing: the commercial applications*, The Knowledge Engineering Review, Vol. 1, No. 3, pp. 11–23.

[60] Jurafsky, D. and H. Martin (2000), *Speech and language processing: An introduction to natural language processing*, Computational linguistics, and speech recognition, New Delhi, India: Pearson Education.

[61] Kamath Aishwarya and Das Rajarshi (2019), *A Survey on Semantic Parsing*, Automated Knowledge Base Construction (AKBC), https://openreview.net/forum?id=HylaEWcTT7

[62] Kawahara, Daisuke and Palmer, Martha (2014), *Single Classifier Approach for Verb Sense Disambiguation based on Generalized Features*, In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 4210–4213.

[63] Kasami Tadao (1965), *An Efficient Recognition and Syntax-Analysis Algorithm for Context-Free Languages*, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, `https://api.semanticscholar.org/CorpusID:61491815`.

[64] Kiperwasser Eliyahu and Goldberg Yoav (2016), *Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations*, Transactions of the Association for Computational Linguistics, Vol. 4, pp 313–327. `https://aclanthology.org/Q16-1023`.

[65] Kiem-Hieu Nguyen (2018), *BKTreebank: Building a Vietnamese Dependency Treebank*, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, European Language Resources Association (ELRA), pp 2164-2168.

[66] Kiet Van Nguyen and Nguyen, Ngan Luu-Thuy (2015), *Error Analysis for Vietnamese Dependency Parsing*, In 2015 Seventh International Conference on Knowledge and Systems Engineering (KSE), pp 79–84, `http://arxiv.org/abs/1911.03724`.

[67] Kingsbury, P., Palmer, M. (2002), *From TreeBank to PropBank*, Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02), pp. 1–8.

[68] Kipper, K., Korhonen, A., Ryant, N., Palmer, M. (2006), *Extending VerbNet with Novel Verb Classes*, Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), pp. 1–8.

[69] K. Gaurav and S. Sebastian and V. Sowmya and R. Siva (2024), *Scope Ambiguities in Large Language Models*, Transactions of the Association for Computational Linguistics, Shanghai, China, Vol 12, pp 738–754.

[70] Klavans, Judith L. and Kan, Min-Yen (1998), *Role of Verbs in Document Analysis*, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, Quebec, Canada, Association for Computational Linguistics, pp 680–686.

[71] Lam Hoang Thanh and Gabriele Picco and Yufang Hou and Young-Suk Lee and Lam M. Nguyen and Dzung T. Phan and Vanessa López and Ramon Fernandez Astudillo (2021), *Ensembling Graph Predictions for AMR Parsing*, Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, pp 8495–8505.

[72] Levin, B. (1993), *English Verb Classes and Alternations: A Preliminary Investigation*, Chicago Press, University.

[73] Li, Bin and Wen, Yuan and Qu, Weiguang and Bu, Lijun and Xue, Nianwen (2016), *Annotating the Little Prince with Chinese AMRs*, Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016), pp. 7–15.

[74] Linh, Ha and Nguyen, Huyen (2019), *A Case Study on Meaning Representation for Vietnamese*, Proceedings of the First International Workshop on Designing Meaning Representations, pp. 148–153.

[75] Linh Ha, Do Dao, Nguyen Huyen, Ngo Quyen, and Doan Dung (2022), *VLSP 2021 - NER Challenge: Named Entity Recognition for Vietnamese*, *VNU Journal of Science: Computer Science and Communication Engineering*, vol. 38, no. 1.

[76] Liang, Percy (2013), *Lambda Dependency-Based Compositional Semantics*, CoRR, Vol. abs/1309.4408.

[77] Liu Yinhan and Ott Myle and Goyal Naman and Du Jingfei and Joshi Mandar and Chen Danqi and Levy Omer and Lewis Mike and Zettlemoyer Luke and Stoyanov Veselin (2019), *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, CoRR, abs/1907.11692, https://openreview.net/forum?id=SyxS0T4tvS.

[78] Loper, Edward and Yi, Szu-ting and Palmer, Martha (2007), *Combining lexical resources: Mapping between PropBank and VerbNet*, In Proceedings of the IWCS-7.

[79] McDonald, Ryan and Nivre, Joakim (2011), *Analyzing and Integrating Dependency Parsers*, *Computational Linguistics*, Vol. 37(1), pp 197–230.

[80] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru (2018), *Model Cards for Model Reporting*, In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'19). Association for Computing Machinery, New York, USA, pp 220–229. https://doi.org/10.1145/3287560.3287596.

[81] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman (2021), *Universal Dependencies*. Computational Linguistics 2021, Vol. 47 (2), pp. 255–308, doi: https://doi.org/10.1162/coli_a_00402

[82] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter (2015), *Efficient and robust automated machine learning*, In Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'15), Vol. 2. MIT Press, Cambridge, MA, USA, pp. 2755–2763.

[83] Marie-Catherine de Marneffe and Christopher D. Manning (2008), *The Stanford Typed Dependencies Representation*, Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation, Coling 2008, pp. 1-8, Coling 2008 Organizing Committee, Manchester, UK.

[84] M Kay. (1986), *Algorithm schemata and data structures in syntactic processing*, Readings in natural language processing, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 35–70.

[85] Ma, Xuezhe and Hu, Zecong and Liu, Jingzhou and Peng, Nanyun and Neubig, Graham and Hovy, Eduard (2018), *Stack-Pointer Networks for Dependency Parsing*, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1403–1414.

[86] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990), *Introduction to WordNet: An On-line Lexical Database*, International Journal of Lexicography, Vol. 3, No. 4, pp. 235–244.

[87] Mikolov Tomas and Chen Kai and Corrado Greg and Dean Jeffrey, *Efficient Estimation of Word Representations in Vector Space*, In 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings.

[88] Mitchell P. Marcus, Mary Ann Marcinkiewicz, Beatrice Santorini (1993), *Building a Large Annotated Corpus of English: The Penn Treebank*, Comput. Linguist., Vol. 19(2), pp. 313–330.

[89] Nivre Joakim, de Marneffe, Marie-Catherine, Ginter Filip, Goldberg Yoav, Hajič Jan, Manning Christopher D., McDonald, Ryan, Petrov, Slav, Pyysalo, Sampo,

Silveira, Natalia, Tsarfaty, Reut, Zeman, Daniel (2016), *Universal Dependencies v1: A Multilingual Treebank Collection*. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, May 2016. European Language Resources Association (ELRA), pp. 1659–1666, `https://aclanthology.org/L16-1262`.

[90] Nivre, Joakim and de Marneffe, Marie-Catherine and Ginter, Filip and Hajič, Jan and Manning, Christopher D. and Pyysalo, Sampo and Schuster, Sebastian and Tyers, Francis and Zeman, Daniel (2020), *Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection*, Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 4034–4043.

[91] Nguyen, Thi Luong, Ha, My Linh, Nguyen, Viet Hung, Nguyen, Thi Minh Huyen, and Le, Hong Phuong (2013), *Building a treebank for Vietnamese dependency parsing*, Proceedings of the 2013 RIVF International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for Future (RIVF), pp. 147-151.

[92] Nguyen, Luong, Hà, Linh, Nguyen, Huyen, and Phuong, Le-Hong (2018), *Using BiLSTM in dependency parsing for Vietnamese*, *Computacion y Sistemas*, 22, pp. 853–862.

[93] Nguyen Luong Tran and Duong Minh Le and Dat Quoc Nguyen (2022), *BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese*, Proceedings of the 23rd Annual Conference of the International Speech Communication Association.

[94] Nguyen, T. M. H., Romary, L., Rossignol, M., and Vu, X. L. (2006), *A lexicon for Vietnamese language processing*, *Language Resources and Evaluation*, 40, pp. 291–309.

[95] Nguyen Huyen T M and Nguyen Hung V and Ngo Quyen T and Vu Luong X and Tran Vu Mai and Ngo Bach X and Le Cuong A (2013), *VLSP Shared task: Sentiment Analysis*, Journal of Computer Science and Cybernetics, Vol. 34, pp. 295–310.

[96] Nguyen, Thai Phuong and Pham, Van-Lam and Nguyen, Hoang-An and Vu, Huy-Hien and Tran, Ngoc-Anh and Truong, Thi-Thu-Ha (2016), *A Two-Phase Approach for Building Vietnamese WordNet*, Proceedings of the 8th Global WordNet Confer-

ence (GWC), edited by Christiane Fellbaum, Piek Vossen, Verginica Barbu Mititelu, and Corina Forascu, pp. 261–266.

[97] OpenAI (2023), *GPT-4 Technical Report*, CoRR, Vol. abs/2303.08774.

[98] Petukhova, Volha and Bunt, Harry (2008), *LIRICS Semantic Role Annotation: Design and Evaluation of a Set of Data Categories*, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco.

[99] Percy Liang, Michael I. Jordan, and Dan Klein (2013), *Learning Dependency-Based Compositional Semantics*, Computational Linguistics, Vol. 39(2), pp. 389–446.

[100] Pennington, Jeffrey and Socher, Richard and Manning, Christopher (2014), *GloVe: Global Vectors for Word Representation*, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), edited by Alessandro Moschitti, Bo Pang, and Walter Daelemans, pp. 1532–1543.

[101] Phuong Thai Nguyen, Luong Vu Xuan, Thi Minh Huyen Nguyen, Van Hiep Nguyen, and Phuong Le-Hong (2009), *Building a Large Syntactically-Annotated Corpus of Vietnamese*, In Proceedings of the Third Linguistic Annotation Workshop (LAW III), Suntec, Singapore, pp. 182–185.

[102] Hong, Phuong, Pham, Hoang, Pham, Khoai, Nguyen, Huyen, Nguyen, Luong, and Nguyen, Hiep (2017), *Vietnamese Semantic Role Labelling*, VNU Journal of Science: Computer Science and Communication Engineering.

[103] Pradet, Quentin and de Chalendar, Gaël and Desormeaux Baguenier, Jeanne (2014), *WoNeF, an improved, expanded and evaluated automatic French translation of WordNet*, Proceedings of the Seventh Global Wordnet Conference (GWC2014), pp. 32–39.

[104] Qi Peng and Zhang Yuhao and Zhang Yuhui and Bolton Jason, and Manning Christopher D. (2020), *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 101-108.

[105] Ralph M. Weischedel and Eduard H. Hovy and Mitchell P. Marcus and Martha Palmer (2017), *OntoNotes : A Large Training Corpus for Enhanced Processing*,

Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation, Vol. 3, No. 3, pp. 3-4.

[106] Reppen, Randi and Ide, Nancy and Suderman, Keith (2005), *American National Corpus (ANC)*, Linguistic Data Consortium.

[107] Rousseeuw, P. J. (1987), *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*, *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53–65.

[108] Shi, Lei and Mihalcea, Rada (2005), *Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing*, Computational Linguistics and Intelligent Text Processing, vol. 3406, pp. 100–111.

[109] Srinivasan Iyer (2019), *Learning to Map Natural Language to General Purpose Source Code*, Thesis of Doctor of Philosophy, University of Washington.

[110] Shui Ruihao and Cao Yixin and Wang Xiang and Chua Tat-Seng (2023), *A Comprehensive Evaluation of Large Language Models on Legal Judgment Prediction*, in Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, pp. 7337–7348.

[111] Strubell, Emma, Verga, Patrick, Belanger, David, and McCallum, Andrew (2018), *Linguistically-informed self-attention for semantic role labeling*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 5027–5038.

[112] Tai, Kai Sheng and Socher, Richard and Manning, Christopher D (2015), *Improved semantic representations from tree-structured long short-term memory networks*, in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Vol. 1, pp. 1556–1566.

[113] Tesnière, Lucien (1959), *Eléments de Syntaxe Structurale*, Klincksieck.

[114] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, Kate Crawford (2018), *Datasheets for Datasets*, Communications of the ACM, 64. 10.1145/3458723.

[115] Tran, Tri, Pham, T., Ngo, Hung, Dien, Dinh, and Collier, Nigel (2007), *Named Entity Recognition in Vietnamese documents*, Progress in Informatics, pp. 5–13.

[116] Van-Nhat Nguyen, Ha-Thanh Nguyen, Dinh-Hieu Vo, and Le-Minh Nguyen (2018), *Relation Extraction in Vietnamese Text via Piecewise Convolution Neural Network with Word-Level Attention*, in *5th NAFOSTED Conference on Information and Computer Science (NICS)*, IEEE, pp. 99–103.

[117] Van-Hai Vu, Quang-Phuoc Nguyen, Kiem-Hieu Nguyen, Joon-Choul Shin, and Cheol-Young Ock (2020), *Korean-Vietnamese Neural Machine Translation with Named Entity Recognition and Part-of-Speech Tags*, IEICE Transactions on Information and Systems, Vol. E103.D, No. 4, pp. 866–873.

[118] Son Vu Xuan, Thanh Vu, Son Tran, and Lili Jiang (2019), *ETNLP: A Visual-Aided Systematic Approach to Select Pre-Trained Embeddings for a Downstream Task*, In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019) Varna, Bulgaria, pp. 1285–1294,

[119] Wang Shan and Bond Francis (2013), *Building the Chinese Open Wordnet (COW): Starting from Core Synsets*, Proceedings of the 11th Workshop on Asian Language Resources, pp. 10–18.

[120] Wein, Shira, Donatelli, Lucia, Ricker, Ethan, Engstrom, Calvin, Nelson, Alex, Harter, Leonie, and Schneider, Nathan (2022), *Spanish Abstract Meaning Representation: Annotation of a General Corpus*, Northern European Journal of Language Technology, Vol. 8.

[121] White, Aaron Steven, Reisinger, Drew, Sakaguchi, Keisuke, Vieira, Tim, Zhang, Sheng, Rudinger, Rachel, Rawlins, Kyle, and Van Durme, Benjamin (2016), *Universal Decompositional Semantics on Universal Dependencies*, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, pp. 1713–1723.

[122] Woods, William A. (1973), *Progress in Natural Language Understanding: An Application to Lunar Geology*, AFIPS National Computer Conference, AFIPS Conference Proceedings, Vol. 42, pp. 441–450.

[123] Xuan, Thao, Kawazoe, Ai, Dien, Dinh, Collier, Nigel, and Tri, Tran (2007), *Construction of a Vietnamese Corpora for Named Entity Recognition*, In Large Scale Semantic Access to Content (Text, Image, Video, and Sound) (RIAO '07), Paris, FRA, pp. 719–724.

[124] Xue Naiwen and Xia, Fei and Chiou, Fu-Dong and Palmer, Marta (2005), *The Penn Chinese TreeBank: Phrase structure annotation of a large corpus*, Natural Language Engineering, Vol. 11, No. 2, pp. 207–238, DOI: 10.1017/S135132490400364X.

[125] Yanshan Wang, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Fei Liu, and Hongfang Liu (2017), *Dependency and AMR Embeddings for Drug-Drug Interaction Extraction from Biomedical Literature*, Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 36–43.

[126] Yang Kaiyu, and Deng Jia (2020), *Strongly Incremental Constituency Parsing with Graph Neural Networks*, Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, pp. 1820–1831.

[127] You Liping and Liu Kaiying (2005), *Building Chinese FrameNet database*, 2005 International Conference on Natural Language Processing and Knowledge Engineering, pp. 301–306.

[128] Y. Shuguang and C. Feipeng and Y. Yiming and Z. Zude (2024), *A Study on Semantic Understanding of Large Language Models from the Perspective of Ambiguity Resolution*, in Proceedings of the 2023 International Joint Conference on Robotics and Artificial Intelligence, Association for Computing Machinery, No. 6, pp 165-170.

[129] Yu Zhang and Houquan Zhou and Zhenghua Li (2020), *Fast and Accurate Neural CRF Constituency Parsing*, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20), pp 4046–4053.

[130] Zhou Junru, and Zhao Hai (2019), "Head-Driven Phrase Structure Grammar Parsing on Penn Treebank," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2396–2408, Florence, Italy.

[131] Zhou, Jie and Xu, Wei (2015), *End-to-end learning of semantic role labeling using recurrent neural networks*, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp 1127–1137.