ĐẠI HỌC QUỐC GIA TP. HCM

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

NGUYỄN TRỌNG CHỈNH

MÔ HÌNH VÀ PHƯƠNG PHÁP LẬP LUẬN ĐỂ TRẢ LỜI CÁC CÂU HỎI "TẠI SAO" DỰA TRÊN CÁCH TIẾP CẬN PHÂN TÍCH DIỄN NGÔN TIẾNG VIỆT

Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số: 62.48.01.01

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

TP. HỒ CHÍ MINH – NĂM 2022

MỤC LỤC

DANH MỤC TỪ VIẾT TẮT

BERT	Bidirectional Encoder Representations from Transformers
CRF	Conditional Random Field
DMN	Dynamic Memory Network
DNN	Deep Neural Network
DRT	Discourse Representation Theory
EDU	Elementary Discourse Unit
FFNN	Feed Forward Neural Network
HMM	Hidden Markov Model
LSA	Latent Semantic Analysis
ME	Maximum Entropy
MRC	Machine Reading Comprehension
MRR	Mean Reciprocal Rank
NLI	Natural Language Inference
NLP	Natural Language Processing
QA	Question Answering
RST	Rhetorical Structure Theory
RTE	Recognizing Textual Entailment
TBL	Transformation Based Learning
WE	Word Embeddings

DANH MỤC CÁC THUẬT NGỮ

TIẾNG ANH	TIẾNG VIỆT	DIỄN GIẢI
Anaphora	Hồi chỉ	Biện pháp nhắc đến những đối tượng đã được nêu trước đó nhưng tránh nêu lại.
Causal relation [59]	Quan hệ nhân–quả [2]	Theo [2], chỉ mối quan hệ "nguyên nhân – kết quả" giữa hai động từ như cặp động từ "give – have" (cho – có).
Cause relation	Quan hệ diễn ngôn chỉ nguyên nhân	Quan hệ giữa hai đơn vị diễn ngôn trong đó một đơn vị diễn ngôn diễn tả nguyên nhân của sự việc được diễn tả trong đơn vị diễn ngôn còn lại.
Context	Khung cảnh, văn cảnh, ngữ cảnh	- Khung cảnh: chỉ môi trường diễn ra sự việc. - Văn cảnh hay ngữ cảnh: chỉ những từ ngữ hoặc những câu xung quanh từ ngữ hoặc câu đang xét.
Defeasible	Không vững (lập luận)	Chỉ khả năng dùng lý lẽ để phản bác một lập luận.
Discourse	Diễn ngôn	Xem **Mục 1.3.2.3**
Discourse marker	Từ ngữ liên kết	Các từ ngữ có chức năng liên kết các mệnh đề, các câu hoặc các đoạn văn trong văn bản, là dấu chỉ của một số loại quan hệ diễn ngôn.
Distributional Semantic	Ngữ nghĩa phân bố	Hướng tiếp cận tính toán ngữ nghĩa dựa vào thực tế sử dụng từ ngữ.
Element discourse unit	Đơn vị diễn ngôn cơ bản	
Entailment	Quan hệ kéo theo	Giả sử A và B là mệnh đề, câu hoặc đoạn văn. Theo Dagan [23], A có quan hệ kéo theo B nếu người đã đọc A sẽ cho rằng B đúng.
Explicit meaning	Nghĩa hiển ngôn [3]	Theo Cao Xuân Hạo [3], *"là nghĩa nguyên văn (gồm nghĩa đen và một số nghĩa bóng quen thuộc) của những từ ngữ có mặt trong*

		câu và nhờ những mối quan hệ cú pháp giữa các từ ấy."
Finite clause	Mệnh đề quan hệ hạn định	Một khái niệm mệnh đề trong tiếng Anh
Implicit meaning	Nghĩa hàm ẩn [3]	Theo Cao Xuân Hạo [3], *"những gì không có sẵn trong nghĩa nguyên văn của các từ ngữ và trong mối quan hệ cú pháp ấy nhưng vẫn thấu đến người nghe thông qua một sự suy diễn."*
Informal argument	Lập luận đời thường (lập luận phi hình thức)	
Internal argument	Lập luận con	Lập luận là tiền đề của một lập luận khác.
Nucleus	Hạt nhân (vai trò trong quan hệ diễn ngôn)	Đơn vị mang nghĩa quan trọng hơn, không thể lược bỏ khi rút gọn.
Presupposition	Tiền giả định	Những khẳng định được ngầm hiểu là đúng.
Real-life arugment	Lập luận đời thường (Lập luận phi hình thức)	
Referent	Sở chỉ	Đối tượng hay sự kiện được nêu trong văn bản.
Result relation	Quan hệ diễn ngôn chỉ kết quả	Quan hệ giữa hai đơn vị diễn ngôn trong đó một đơn vị diễn ngôn diễn tả kết quả của một sự việc được diễn tả trong đơn vị diễn ngôn còn lại.
Rhetorical structure	Cấu trúc tu từ	Cấu trúc giúp nhấn mạnh ý cần diễn đạt.
Satellite	Vệ tinh (vai trò trong quan hệ diễn ngôn)	Đơn vị mang nghĩa ít quan trọng hơn, có thể lược bỏ khi rút gọn văn bản.
Validity	Tính hiệu lực (của lập luận)	Cho biết phép kéo theo từ tiền đề đến kết luận là đúng
Word embedding	Vector từ	Nghĩa của từ được biểu diễn dưới dạng vector

MỞ ĐẦU

Lý do lựa chọn đề tài

Luận án chọn đề tài vì bốn lý do sau:

- Các kết quả nghiên cứu đã được công bố về phương pháp trả lời câu hỏi "TẠI SAO" (Why-question) chưa nhiều và hiệu quả của các phương pháp còn chưa cao.
- Vấn đề phân tích diễn ngôn chưa thể hiện rõ ràng trong các nghiên cứu phương pháp tìm câu trả lời cho câu hỏi "TẠI SAO".
- Cách tiếp cận lập luận dựa trên cấu trúc diễn ngôn để trả lời câu hỏi "TẠI SAO" chưa được nghiên cứu trong khi lập luận và cấu trúc diễn ngôn là phương tiện để trình bày lý lẽ và giải thích.
- Cách tiếp cận lập luận dựa trên cấu trúc diễn ngôn có thể chỉ ra quá trình lập luận để tìm câu trả lời.

Mục đích của luận án

Mục đích của luận án là nghiên cứu phương pháp lập luận và phương pháp phân tích diễn ngôn làm cơ sở để xác lập cơ chế tìm câu trả lời trong văn bản tiếng Việt cho các câu hỏi "TẠI SAO". Cơ chế này có đặc điểm hai đặc điểm:

- Quá trình tìm câu trả lời được thể hiện rõ ràng, có thể theo dõi được.
- Phù hợp với cách suy luận để tìm câu trả lời của người Việt vì được xây dựng từ diễn ngôn tiếng Việt và lập luận.

Nội dung nghiên cứu

Để đạt được mục đích nghiên cứu, các nội dung cần được nghiên cứu trong luận án như sau:

- Tổng quan về hỏi-đáp tự động và các nghiên cứu về câu hỏi "TẠI SAO".

- Mô hình và phương pháp phân tích một số quan hệ diễn ngôn cấp độ câu và liên câu trong tiếng Việt.
- Phương pháp lập luận trên biểu diễn văn bản.
- Phương pháp xác định câu trả lời cho các câu hỏi "TẠI SAO" dựa trên lập luận và phân tích diễn ngôn.
- Mô hình lập luận để trả lời câu hỏi "TẠI SAO" dựa trên cách tiếp cận phân tích diễn ngôn cho văn bản tiếng Việt.

Đối tượng nghiên cứu

Từ mục đích của luận án, đối tượng nghiên cứu được xác định gồm:

- Mô hình diễn ngôn áp dụng để biểu diễn diễn ngôn cho văn bản tiếng Việt.
- Đơn vị diễn ngôn trong văn bản tiếng Việt.
- Quan hệ diễn ngôn trong văn bản tiếng Việt.
- Lập luận theo cơ chế loại suy.
- Nhận dạng lập luận và tạo lập luận theo cơ chế loại suy.

Phạm vi nghiên cứu

Luận án được nghiên cứu trong phạm vi như sau:

- Phạm vi áp dụng là dạng bài toán là tìm câu trả lời cho câu hỏi “TẠI SAO” trong một văn bản ngắn.
- Phân tích quan hệ diễn ngôn trong luận án chỉ thực hiện ở cấp độ câu và liên câu cho một số quan hệ diễn ngôn được chọn, phục vụ cho việc xác định câu trả lời cho câu hỏi “TẠI SAO” của luận án.
- Lập luận loại suy được thực hiện theo theo lược đồ loại suy của Juthe.
- Tính toán ngữ nghĩa được thực hiện với nghĩa thông thường của từ ngữ, không tính toán nghĩa hàm ẩn.

Ý nghĩa khoa học và thực tiễn của đề tài

Luận án có bốn đóng góp chính sau:

- Phân tích diễn ngôn ở cấp độ câu và liên câu theo một số quan hệ được chọn. Các kết quả nghiên cứu được công bố trong các công trình [CT.3] và [CT.6] và có liên quan đến công trình [CT.1]
- Phương pháp lập luận trên biểu diễn văn bản tiếng Việt theo lược đồ lập luận loại suy. Các kết quả nghiên cứu được công bố trong các công trình [CT.4] và [CT.8]
- Phương pháp lập luận để trả lời các câu hỏi "TẠI SAO" dựa trên cách tiếp cận phân tích diễn ngôn tiếng Việt. Phương pháp này có một ưu điểm nổi bật là có thể tìm được các ý trong một câu trả lời có nhiều hơn một ý; trong đó, các ý này là các chuỗi không liên tục trong văn bản. Các kết quả nghiên cứu được công bố trong các công trình nghiên cứu [CT.2], [CT.5] và [CT.7]
- Mô hình lập luận để trả lời các câu hỏi "TẠI SAO" dựa trên cấu trúc diễn ngôn tiếng Việt được công bố trong công trình nghiên cứu [CT.7]

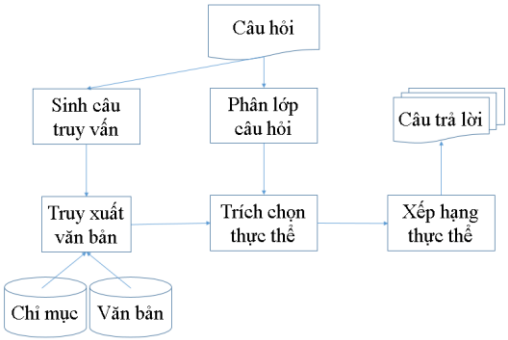
Bên cạnh đó, kết quả xây dựng bộ ngữ liệu gán nhãn đơn vị diễn ngôn cơ bản dựa trên ngữ liệu phân tích cú pháp cấu trúc ngữ đoạn tiếng Việt NIIVTB và bộ ngữ liệu VnNewsNLI tiếng Việt của luận án cũng hỗ trợ cho các nghiên cứu về phân tích diễn ngôn tiếng Việt và suy luận trên biểu diễn văn bản tiếng Việt.

CHƯƠNG 1. TỔNG QUAN

1.1 HỎI-ĐÁP TỰ ĐỘNG

1.1.1 Nguồn gốc bài toán

Hỏi-đáp tự động (question answering) là một nhánh nghiên cứu trong truy xuất thông tin (Information Retrieval). Theo khảo sát của tác giả Kolomiyets [47], nhiều phương pháp giải quyết bài toán hỏi-đáp được nghiên cứu dựa trên cơ sở của truy xuất thông tin kết hợp với phương pháp phân loại câu hỏi (Text Classification) [54, 63] và phương pháp trích chọn thực thể trong văn bản (Named Entity Recognition) [65, 104]. Theo hướng này, phương pháp chung để giải quyết bài toán hỏi-đáp tự động có thể được minh họa bằng **Hình 1.1**



Hình 1.1 Sơ đồ các bước xử lý trong phương pháp hỏi-đáp tự động

Phương pháp chung để giải quyết bài toán hỏi-đáp tự động trong **Hình 1.1** gồm các bước chính:

1) Sinh câu truy vấn. Tạo câu truy vấn từ câu hỏi của người sử dụng.
2) Truy xuất văn bản. Chọn danh sách tài liệu có liên quan đến câu truy vấn.

3) Phân lớp câu hỏi. Câu hỏi được phân lớp theo nội dung cần có của câu trả lời.

4) Trích chọn thực thể. Chọn các thực thể phù hợp với nội dung cần có của câu trả lời.

5) Xếp hạng thực thể. Đánh giá mức độ phù hợp giữa câu trả lời tiềm năng với câu hỏi và chọn câu trả lời tiềm năng.

Qua các nghiên cứu về phân lớp câu hỏi của [39], [54] và [35], một danh mục các phân lớp câu hỏi được xây dựng để phục vụ cho việc nghiên cứu phương pháp tìm câu trả lời cho các dạng câu hỏi.

Dựa vào các lớp câu hỏi, các nghiên cứu về hỏi-đáp tự động được chia thành hai nhóm là Factoid và Non-factoid. Trong nhóm Non-factoid, câu hỏi "TẠI SAO" (Why-question) có một điểm khó là câu trả lời có ngữ nghĩa không tương tự như câu hỏi mà nó là lý do dẫn đến sự việc được nêu trong câu hỏi.

1.1.2 Bài toán trả lời câu hỏi "TẠI SAO"

Bài toán trả lời câu hỏi "TẠI SAO" có hai dạng gồm dạng đầy đủ và dạng rút gọn. Dạng đầy đủ của bài toán trả lời câu hỏi "TẠI SAO" được phát biểu như sau.

Cho $\boldsymbol{Docs} = \{\boldsymbol{d}_i | i = \overline{1,n}\}$ là một tập hợp các tài liệu văn bản, $\boldsymbol{q}$ là một câu hỏi "TẠI SAO". Tìm $\boldsymbol{a}$ là một tập các chuỗi ký tự có trong các tài liệu $\boldsymbol{d}_j \in \boldsymbol{Docs}$ sao cho nội dung của $\boldsymbol{a}$ là lý do giải thích cho nội dung của $\boldsymbol{q}$. Khi đó $\boldsymbol{a}$ là câu trả lời cho câu hỏi $\boldsymbol{q}$.

Dạng rút gọn của bài toán trả lời câu hỏi "TẠI SAO" có thể phát biểu như bài toán đọc hiểu văn bản được Ellen Riloff và Michael Thelen đề xuất [78] như sau.

Cho $\boldsymbol{d}$ là một tài liệu văn bản, $\boldsymbol{q}$ là một câu hỏi "TẠI SAO". Tìm $\boldsymbol{a}$ là một chuỗi ký tự trong của một tài liệu $\boldsymbol{d}$ sao cho nội dung của $\boldsymbol{a}$ giải thích cho nội dung được hỏi trong câu hỏi $\boldsymbol{q}$. Khi đó $\boldsymbol{a}$ là câu trả lời cho câu hỏi $\boldsymbol{q}$.

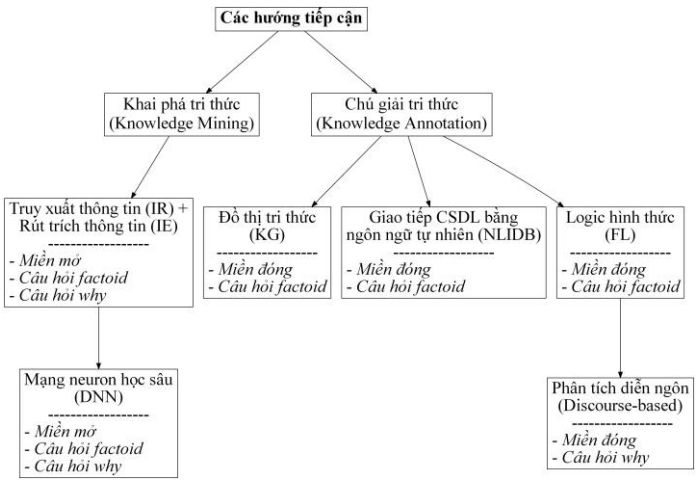
Phát biểu bài toán trả lời câu hỏi "TẠI SAO" trong cả hai dạng như trên sẽ không phù hợp trong trường hợp câu trả lời có nhiều hơn một ý. Vì thế, luận án sẽ xem vấn đề tìm câu trả lời có nhiều ý như là một ưu điểm của phương pháp tìm câu trả lời của luận án so với các kết quả nghiên cứu đã được công bố.

1.1.3 Đánh giá phương pháp trả lời câu hỏi

Phương pháp đánh giá theo thực nghiệm sử dụng một bộ ngữ liệu $Gold = \{\langle d_i, q_i, a_i \rangle | i = \overline{1, n}\}$. Kết quả trả lời các câu hỏi q_i dựa vào các tài liệu d_i của hệ thống sẽ được so sánh với câu trả lời a_i và đánh giá theo các độ đo MRR, độ phủ, độ chính xác và F_1.

1.2 TỔNG QUAN TÌNH HÌNH NGHIÊN CỨU

Theo khảo sát [27, 47, 63], có hai hướng tiếp cận là khai phá tri thức và chú giải tri thức như được trình bày ở Hình 1.2.



Hình 1.2 Các hướng tiếp cận giải quyết bài toán hỏi-đáp

1.2.1 Hướng tiếp cận chú giải tri thức

Hướng chú giải tri thức tập trung vào nghiên cứu biểu diễn (representation) tri thức của bài toán, xây dựng tri thức và nghiên cứu các quy tắc tính toán trên các biểu diễn đó. Bên cạnh đó, một số nghiên cứu về phương pháp trả lời câu hỏi theo cách tiếp cận biểu diễn tri thức theo logic hình thức, cụ thể là logic vị từ bậc một (first-order logic) [9, 10, 45] mặc dù chưa được triển khai thành hệ thống hỏi-đáp nhưng cũng thể hiện tính khả thi của cách tiếp cận này. Chẳng hạn phương pháp của Delmonte [24] có thể tìm câu trả lời cho câu hỏi "TẠI SAO" dựa trên một số quan hệ diễn ngôn.

1.2.2 Hướng tiếp cận khai phá tri thức

Theo Mishra và cộng sự [63], Dimitrakis và cộng sự [27] thì các phương pháp trả lời câu hỏi theo hướng tiếp cận khai phá tri thức trong giai đoạn từ năm 2002 đến 2014 được phát triển dựa trên sự kết hợp truy xuất thông tin [47] với trích chọn thông tin.

Từ năm 2014, nhiều kết quả nghiên cứu từ mạng nơron học sâu, như Dynamic Memory Network [83] và BERT [26], được áp dụng trong các nghiên cứu xác định câu trả lời trong bài toán đọc hiểu văn bản (machine reading comprehension). Các kết quả nghiên cứu này được sử dụng để xây dựng các hệ thống end-to-end [22] có hiệu quả cao khi trả lời các câu hỏi factoid.

1.2.3 Các phương pháp giải quyết vấn đề trả lời câu hỏi "TẠI SAO"

1.2.3.1 Cách tiếp cận phân tích diễn ngôn

Phương pháp của Delmonte [24] là phương pháp duy nhất theo cách tiếp cận này. Theo đó, đoạn văn bản được phân tích thành các biểu thức logic vị từ bậc một và các quan hệ diễn ngôn trong bốn nhóm do Delmonte đề xuất [24] là Cause-Result, Rationale-Effect, Purpose-Outcome, Circumstance-Outcome và Means-Outcome.

1.2.3.2 Cách tiếp cận kết hợp IR và IE

Trong cách tiếp cận này, các phương pháp trả lời câu hỏi "TẠI SAO" dùng một mô hình IR để chọn các đoạn văn. Sau đó dùng một mô hình IE để trích các câu trả lời trong các đoạn văn bản. Các phương pháp này được tổng hợp và trình bày trong Bảng 1.2.

Bảng 1.2 Các phương pháp trả lời câu hỏi "TẠI SAO" theo cách tiếp cận kết hợp IR và IE

Tác giả	**Năm**	**Phương pháp**	**Ngữ liệu**	**Kết quả**
Verberne	2006-2010	IR + phân lớp quan hệ RST	186 câu hỏi tại sao được tác giả chọn từ ngữ liệu **INEX**	MRR@150 = 0,34
Higashinaka and Isozaki	2008	IR + phân lớp quan hệ nguyên nhân – kết quả với phương pháp phân lớp SVM	Ngữ liệu tiếng Nhật do tác giả tự xây dựng	MRR@20 = 0,339
Azmi	2016-2017	IR + phân tích diễn ngôn theo RST	Ngữ liệu tiếng Ả rập của tác giả	ACC = 71%
Oh et. al.	2013	IR + Rút trích các mệnh đề chứa quan hệ nguyên nhân – kết quả giữa các ngữ đoạn, sử dụng mô hình CRF.	Ngữ liệu tiếng Nhật **WhySet** do tác giả xây dựng	P@1= 41,8%
	2016	IR + Rút trích các mệnh đề chứa quan hệ nguyên nhân – kết quả giữa các ngữ đoạn, sử dụng mô hình CRF. Tăng cường hiệu quả bằng việc thêm ngữ liệu huấn luyện.	Ngữ liệu tiếng Nhật **WhySet**	P@1 = 50%
	2017	IR + Rút trích các mệnh đề chứa quan hệ nguyên nhân – kết quả giữa các ngữ đoạn, sử dụng mô hình CRF. Sử dụng mạng nơron CNN để chọn câu trả lời	Ngữ liệu tiếng Nhật **WhySet**	P@1 = 54%
	2019	IR + dùng mạng nơron tương tự GAN (GAN –	Ngữ liệu tiếng Nhật **WhySet**	P@1 = 54,8%

		generative adversarial network)	**Quasar-T**[1]	EM = 43,2% F_1 = 49,7%
			SearchQA	EM = 59,6% F_1 = 65,3%
			TriviaQA	EM = 49,6% F_1 = 54,8%

1.2.3.3 Cách tiếp cận dùng mạng nơron học sâu

Cách tiếp cận dùng mạng nơron học sâu phù hợp với bài toán đọc hiểu văn bản. Bài toán đọc hiểu văn bản không cần bước truy xuất văn bản mà chỉ tập trung vào bước trích chọn câu trả lời. Kết quả trích chọn câu trả lời của các mô hình r-net+ (ensemble) [99], SLQA+ (ensemble) [98], Match-LSTM (boundary+ensemble) [97] và BERT (ensemble) trên tập development của ngữ liệu thử nghiệm SQuAD v1.1 [77] đã được công bố tại website cung cấp ngữ liệu SQuAD[2], được trích lại trong Bảng 1.3. Từ Bảng 1.3 cho thấy, mặc dù kết quả trích chọn câu trả lời cho tất cả dạng câu hỏi có độ đo F_1 khá cao (cao nhất là 93,16% theo mô hình BERT) nhưng kết quả trích chọn câu trả lời cho câu hỏi "TẠI SAO" lại có độ đo F_1 thấp hơn nhiều (cao nhất là 69,66% theo mô hình BERT).

Bảng 1.3 Kết quả trả lời các dạng câu hỏi và dạng câu hỏi tại sao trên tập development của ngữ liệu SQuAD v1.1

Hệ thống	**F_1**	
	Tất cả câu hỏi	**Câu hỏi "TẠI SAO"**
r-net+ (ensemble)	88,49%	66,90%
SLQA+ (ensemble)	88,61%	65,69%
Match-LSTM (Boundary+Ensemble)	77,02%	56,95%
BERT (ensemble)	**93,16%**	**69,66%**

[1] https://github.com/bdhingra/quasar

[2] https://rajpurkar.github.io/SQuAD-explorer/

1.2.4 Nhận xét các phương pháp trả lời câu hỏi "TẠI SAO"

Nhận xét thứ nhất, các phương pháp trả lời câu hỏi "TẠI SAO" sử dụng đặc trưng liên quan đến cấu trúc diễn ngôn gồm từ ngữ liên kết và quan hệ diễn ngôn nhưng chưa giải quyết bài toán phân tích diễn ngôn. Kết quả phân tích diễn ngôn còn chưa đảm bảo tính hệ thống.

Nhận xét thứ hai, các phương pháp theo cách tiếp cận mạng nơron học sâu sử dụng các mẫu hỏi-đáp để huấn luyện các mô hình học máy nhằm dự đoán vị trí câu trả lời trong một đoạn văn bản. Quá trình tính toán của các mô hình học máy khi tìm câu trả lời không cho thấy được rằng câu trả lời đã được xác định như thế nào.

Nhận xét thứ ba, các phương pháp xác định câu trả lời sử dụng một mô hình phân lớp để nhận dạng một cặp các câu có cấu trúc nguyên nhân – kết quả cho thấy mô hình phân lớp có thể dùng trong bài toán suy luận và suy luận là một cơ chế cần phải sử dụng để tìm câu trả lời cho câu hỏi "TẠI SAO".

1.3 CÁC VẤN ĐỀ LIÊN QUAN

1.3.1 Lập luận

Khái niệm lập luận được dùng với nhiều nghĩa. Với nghĩa danh từ, khái niệm lập luận được dùng để chỉ một cấu trúc được sử dụng trong ngôn ngữ tự nhiên hay logic hình thức. Với nghĩa động từ, khái niệm lập luận, hay suy luận, được dùng để chỉ quá trình tạo ra một lập luận.

Lập luận được phân chia thành hai loại lập luận hình thức và lập luận đời thường. Theo Walton và cộng sự [96], mối quan hệ tuần tự, mối tương quan giữa các sự kiện, mối quan hệ giữa nguyên nhân và sự ảnh hưởng của nó cũng có thể tạo thành một lập luận đời thường. Lập luận cần tuân theo luật suy diễn hoặc lược đồ lập luận để đảm bảo tính hiệu lực hoặc tính vững chắc của lập luận.

Johnson và cộng sự [41] đã chỉ ra rằng, có những dấu hiệu, chẳng hạn từ ngữ và cấu trúc câu, được sử dụng trong ngôn ngữ tự nhiên để giúp cho việc nhận dạng lập luận trong hoạt động ngôn ngữ được trở nên dễ dàng hơn. Bên cạnh các lập luận, lời giải thích (explanation) cũng trình bày các chứng cứ hoặc lý do của một tuyên bố nhưng không xét đến tính hiệu lực. Mặc dù không phải là lập luận, lời giải thích có hình thức rất giống lập luận.

1.3.2 Diễn ngôn

Câu trả lời của câu hỏi "TẠI SAO" không chỉ có trong lập luận mà còn có trong các câu hoặc đoạn văn diễn tả mục đích của hành vi (purpose), động lực của hành vi (motivation) và hoàn cảnh diễn ra hành vi đó (circumstance) theo nghiên cứu của Verberne [90-93] và Delmonte [24]. Điều này cũng phù hợp với nhận định của Walton và cộng sự [96]. Các câu hoặc đoạn văn này có thể là những lời giải thích chứ chưa phải là lập luận nên không thể nhận dạng bằng cách lập luận mà phải bằng cách phân tích diễn ngôn.

Khái niệm diễn ngôn được xây dựng trên các khái niệm câu, phát ngôn và được phát biểu theo tác giả Đỗ Hữu Châu [1]: *"Diễn ngôn là một quá trình sản sinh ra và liên kết các phát ngôn thành một chỉnh thể. Nó cũng là tên gọi của cái sản phẩm ngôn từ do quá trình đó tạo nên."* Để biểu diễn cấu trúc diễn ngôn trên máy tính, có hai lý thuyết nổi bật được nghiên cứu và áp dụng trên máy tính là Discourse Representation Theory (DRT) của Kamp[45] và Rhetorical Structure Theory (RST) của Mann và Thompson[55].

Có những lập luận đời thường, là những lập luận không vững (defeasible argument) do được xác lập theo cơ chế quy nạp [44, 95, 96], được sử dụng để thuyết phục người tiếp nhận nên chúng được trình bày theo những cấu trúc diễn ngôn phù hợp để nâng cao hiệu quả. Nghiên cứu của Cabrio (2013) [14] cho thấy rằng tồn tại mối liên hệ giữa lập luận và cấu trúc diễn ngôn cho dù mối liên hệ này không phải lúc nào cũng rõ ràng [81]. Hơn nữa, lập luận đời thường sử dụng ngôn

ngữ tự nhiên để truyền tải nên người đưa ra lập luận thường dùng các dấu chỉ lập luận (argumentation indicators) [87] là những từ ngữ liên kết để người tiếp nhận dễ dàng nhận biết. Vì thế, phân tích diễn ngôn cũng liên quan đến việc xác định các lý do từ những lời giải thích hoặc từ các lập luận.

1.4 HƯỚNG TIẾP CẬN CỦA LUẬN ÁN

Hướng tiếp cận của luận án là hướng chú giải tri thức, nghiên cứu triển khai các khái niệm trong ngôn ngữ học và logic trên máy tính liên quan đến vấn đề trả lời câu hỏi "TẠI SAO" và tính toán dựa trên các khái niệm này để xác định câu trả lời trong một văn bản. Hướng tiếp cận của luận án cần giải quyết hai bài toán trọng yếu.

1.4.1 Tính toán mức độ thỏa lược đồ lập luận loại suy

Giải quyết vấn đề nhận dạng lập luận cần phải tính toán mức độ thỏa các lược đồ lập luận của một cặp tiền đề và kết luận. Nếu mức độ thỏa một trong các lược đồ lập luận của một cặp tiền đề và kết luận đạt một ngưỡng nào đó thì có thể kết luận cặp tiền đề và kết luận đó là một lập luận.

1.4.2 Nhận dạng quan hệ diễn ngôn

Giải quyết vấn đề nhận dạng lời giải thích và lập luận trong văn bản cần phải phân tích được cấu trúc diễn ngôn của văn bản. Vấn đề phân tích cấu trúc diễn ngôn của văn bản tiếng Việt là một vấn đề lớn nên luận án chỉ tập trung nhận dạng một số quan hệ diễn ngôn gồm các quan hệ chỉ nguyên nhân (Cause), kết quả (Result), mục đích (Purpose), động lực (Motivation) và khung cảnh (Circumstance). Các quan hệ này đã được Verberne [89-92] và Delmonte [25] sử dụng để xác định câu trả lời cho câu hỏi "TẠI SAO".

1.5 CẤU TRÚC CỦA LUẬN ÁN

Luận án trình bày các kết quả nghiên cứu mô hình và phương pháp lập luận để trả lời câu hỏi "TẠI SAO" dựa trên cách tiếp cận phân tích diễn ngôn tiếng Việt qua năm chương với các nội dung như sau:

- Chương 1 giới thiệu bài toán hỏi-đáp và bài toán trả lời câu hỏi "TẠI SAO", phương pháp đánh giá kết quả của một hệ thống hỏi-đáp tự động, tổng quan tình hình nghiên cứu về hỏi-đáp với câu hỏi "TẠI SAO" và các hướng tiếp cận để giải quyết bài toán hỏi-đáp có thể áp dụng để giải quyết cho câu hỏi "TẠI SAO". Từ tình hình nghiên cứu này lựa chọn hướng tiếp cận và các bài toán trọng yếu cần giải quyết trong luận án.
- Chương 2 trình bày cơ sở lý thuyết gồm RST, lập luận loại suy, kiến trúc mạng BERT. Từ các cơ sở lý thuyết này, luận án đề xuất các khái niệm quan hệ lý do, khái niệm câu trả lời của câu hỏi "TẠI SAO", khái niệm độ thuyết phục của lập luận làm cơ sở cho các nghiên cứu của luận án.
- Chương 3 trình bày phương pháp phân tích diễn ngôn tiếng Việt ở cấp độ câu và liên câu cho văn bản tiếng Việt theo một số quan hệ diễn ngôn có thể chứa câu trả lời cho câu hỏi "TẠI SAO".
- Chương 4 trình bày phương pháp lập luận trên biểu diễn văn bản tiếng Việt trong đó tiền đề và kết luận được giới hạn trong phạm vi một câu.
- Chương 5 trình bày mô hình hệ thống hỏi-đáp với câu hỏi "TẠI SAO" tiếng Việt của luận án. Kết quả đánh giá mô hình của luận án được so sánh với một số phương pháp trả lời câu hỏi "TẠI SAO" đã được công bố.
- Phần kết luận và hướng phát triển trình bày tóm tắt những kết quả đạt được của luận án và các công việc cần nghiên cứu để cải tiến và phát triển các kết quả nghiên cứu của luận án.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1 RHETORICAL STRUCTURE THEORY

Theo Mann và Thompson [54], văn bản là một tập các đơn vị diễn ngôn liên kết với nhau theo trình tự nhất định bởi các quan hệ diễn ngôn để diễn đạt ý của người viết. Để phân tích diễn ngôn của văn bản, cần phải xác định các đơn vị diễn ngôn và các quan hệ diễn ngôn giữa chúng trong văn bản. Kết quả phân tích diễn ngôn của văn bản là một cấu trúc cây.

2.1.1 Đơn vị diễn ngôn cơ bản

Theo Mann và Thompson [54], đơn vị diễn ngôn cơ bản (EDU – Elementary Discourse Unit) phải là một đơn vị văn bản có tính độc lập. Nó là những mệnh đề trừ các mệnh đệ quan hệ hoặc mệnh để chủ ngữ hoặc mệnh đề bổ ngữ. Xét trong một quan hệ diễn ngôn thì các đơn vị diễn ngôn được chia làm hai loại là hạt nhân (nucleus) và vệ tinh (satelite) dựa vào vai trò của nó.

2.1.2 Các quan hệ diễn ngôn

Các quan hệ diễn ngôn trong RST có thể được định nghĩa tùy thẹo theo quan điểm phân tích. Vì thế, số lượng và ý nghĩa của của các quan hệ diễn ngôn không thống nhất ở các công trình nghiên cứu khác nhau.

2.1.3 Nguyên tắc phân tích diễn ngôn văn bản theo cấu trúc RST

Theo Mann và Thompson[54], quá trình phân tích cấu trúc của một văn bản là một tập hợp các lần áp dụng các lược đồ của các quan hệ diễn ngôn trên các đơn vị diễn ngôn. Ví dụ minh họa quá trình phân tích cấu trúc của văn bản được trình bày ở **Phụ lục A.3**.

2.1.4 Phương pháp phân đoạn EDU

Phân đoạn EDU là bài toán đầu tiên trong phân tích diễn ngôn theo RST nhằm tách văn bản thành các EDU. Kết quả phân đoạn EDU ảnh hưởng đến kết quả đánh giá câu trả lời của câu hỏi "TẠI SAO" theo độ đo F_1.

Theo cách tiếp cận dựa trên luật, các chương trình phân đoạn EDU [50, 56] có độ F_1 vào khoảng 86%. Theo cách tiếp cận học máy, các chương trình phân đoạn EDU được xây dựng bằng cách áp dụng phương pháp gán nhãn dữ liệu chuỗi (sequential labeling) như HILDA [36], ToNy [63] và GumDrop [107] có hiệu quả cao với F_1 trên 95%.

2.1.5 Phương pháp xác định quan hệ diễn ngôn

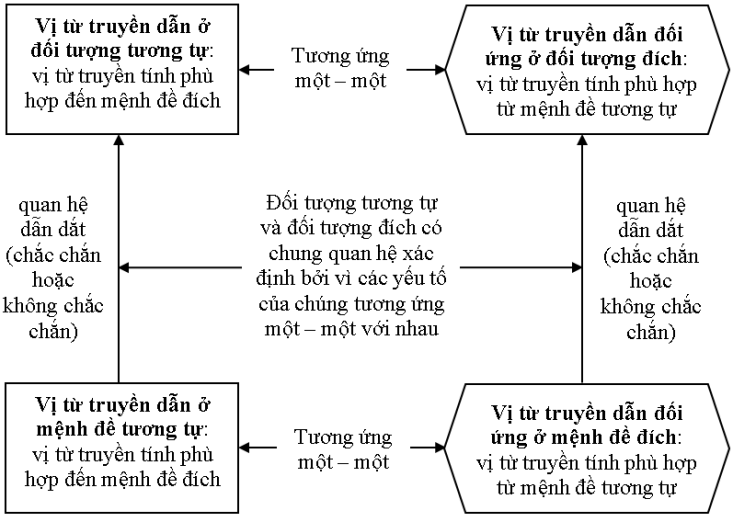
Bài toán xác định quan hệ diễn ngôn tương tự như bài toán phân tích cú pháp phụ thuộc trong đó EDU và nhãn quan hệ diễn ngôn tương ứng với từ vựng và nhãn quan hệ phụ thuộc.

Bài toán xác định quan hệ diễn ngôn là một bài toán khó. Kết quả đánh giá phương pháp của Feng (2012) [30] đạt độ chính xác (accuracy) là 65,3%, của Ji (2014) [40] đạt độ đo F_1 là 61,75%, của Joty (2015) [42] đạt độ đo F_1 là 67,5% trên bộ ngữ liệu RST-DT [16]. Kết quả mới nhất của Zhang (2021) [109] tính toán trên cấu trúc cây diễn ngôn của văn bản có nhãn quan hệ (LAS) đạt độ đo F_1 là 57,6%.

2.2 LẬP LUẬN LOẠI SUY

Lập luận loại suy được chọn làm cơ sở để trả lời các câu hỏi "TẠI SAO" trong luận án bởi vì tính phổ biến của nó trong các suy luận đời thường. Nghiên cứu về lập luận loại suy của Juthe [43] đã chỉ ra một cơ chế chung cho tất cả các dạng lập luận loại suy nên luận án giải quyết vấn đề lập luận theo cơ chế loại suy của Juthe [43].

Lập luận loại suy (argumentation by analogy, argument by analogy hay analogical argument) được định nghĩa theo Stanford Encyclopedia of Philosophy [8] là "*An* ***analogical argument*** *is an explicit representation of a form of analogical reasoning that cites accepted similarities between two systems to support the conclusion that some further similarity exists*". Theo Juthe [43], một lập luận loại suy có các thành phần gồm đối tượng loại suy và đối tượng đích, vị từ truyền dẫn và sự so sánh. Lược đồ cơ bản của lập luận loại suy được Juthe nghiên cứu như Hình 2.3.



Hình 2.3 Lược đồ cơ bản của lập luận loại suy theo Juthe[43]

Đối tượng loại suy (Analogue – A) là đối tượng được dùng để so sánh và lấy ra những đặc điểm cần hỗ trợ cho kết luận.

Vị từ truyền dẫn (Assigned-Predicate – AP) có chức năng tương tự như quan hệ kéo theo trong suy luận diễn dịch. Juthe sử dụng khái niệm vị từ truyền dẫn, ký hiệu là $\prec$ thay cho quan hệ kéo theo.

Sự so sánh (Comparison) là việc đối chiếu sự tương ứng một – một giữa các yếu tố trong mệnh đề loại suy với mệnh đề đích theo tiêu chuẩn của sự tương tự mà Juthe đã đề xuất [43]

Lược đồ lập luận loại suy của Juthe [43] phù hợp với ngôn ngữ tự nhiên vì có thể xác định sự vững chắc của một lập luận một cách trực tiếp dựa vào sự tương tự giữa nó và những lập luận vững chắc cho trước.

2.3 DISTRIBUTIONAL SEMANTIC

Để giải quyết bài toán tính toán sự tương đồng của hai ngữ đoạn, hai câu hay hai đoạn văn, luận án chọn cách tiếp cận ngữ nghĩa học tính toán (computational semantics) trên cơ sở distributional semantic [11]. Distributional semantic là một hướng tiếp cận tính toán ngữ nghĩa khác với ngữ nghĩa hình thức. Distributional semantic [19] sử dụng các vector để biểu diễn nghĩa của từ và nghĩa của các đơn vị cao hơn từ. Các quy tắc tính toán ngữ nghĩa của đơn vị văn bản lớn hơn được thể hiện qua các công thức tính toán các vector nghĩa của các đơn vị văn bản nhỏ hơn.

Ngữ nghĩa cũng được biểu diễn bằng một vector có cùng số chiều như nghĩa của từ. Vì thế, việc tính toán ngữ nghĩa của một ngữ đoạn cũng tương tự như việc tính toán ngữ nghĩa của một câu, một đoạn văn hay thậm chí của cả văn bản. Việc tính toán ngữ nghĩa có thể được thực hiện bởi một mạng nơron. Sự kết hợp nghĩa của từ khi tính toán nghĩa của một đơn vị lớn hơn được xác định thông qua các tham số trong mạng nơron.

Kiến trúc mạng BERT [26] được giới thiệu chính thức vào năm 2019 với khả năng giải quyết nhiều bài toán trong NLP với kết quả tân tiến nhất (state-of-the-art) cho thấy sự phù hợp của BERT trong việc tính toán ngữ nghĩa chuỗi theo hướng tiếp cận distributional semantics. Luận án sử dụng kiến trúc BERT để giải quyết bài toán tính toán sự tương tự theo lược đồ loại suy của Juthe [43] và bài toán phân đoạn EDU.

2.4 ĐỀ XUẤT CÁC KHÁI NIỆM

2.4.1 Khái niệm chuỗi

Để thuận tiện trong việc trình bày và để thống nhất với từ "chuỗi" được sử dụng trong phát biểu bài toán đọc hiểu văn bản của Ellen Riloff và Michael Thelen đề xuất [78], chuỗi sử dụng trong luận án được phát biểu theo Khái niệm 2.1.

Khái niệm 2.1 Chuỗi

Một chuỗi là một ngữ đoạn, một mệnh đề, một câu hoặc một nhóm câu liên tiếp trong văn bản. Chuỗi là đối tượng cần trích chọn làm câu trả lời trong bài toán hỏi-đáp.

2.4.2 Các khái niệm liên quan đến lập luận

Luận án đề xuất các khái niệm làm nền tảng để triển khai lược đồ lập luận loại suy của Juthe trên máy tính gồm:

Khái niệm 2.2 Độ tương tự của hai lập luận: là khái niệm được luận án đề xuất làm cơ sở để xác định một cặp tiền đề và kết luận có thỏa lược đồ lập luận loại suy với một lập luận cho trước hay không.

Khái niệm 2.3 Độ thuyết phục của một lập luận: là khái niệm được luận án đề xuất để tính xác định một cặp tiền đề và kết luận có là một lập luận hay không theo cơ chế loại suy.

2.4.3 Các khái niệm cơ bản của bài toán trả lời câu hỏi "TẠI SAO"

Khái niệm 2.4 Quan hệ lý do giữa hai chuỗi: là khái niệm được luận án đề xuất để chỉ các quan hệ diễn ngôn và quan hệ kéo theo được sử dụng để tìm câu trả lời cho câu hỏi "TẠI SAO"

Khái niệm 2.5 Câu trả lời của câu hỏi "TẠI SAO": là khái niệm được luận án đề xuất, làm cơ sở cho phương pháp lập luận để trả lời câu hỏi "TẠI SAO" dựa trên kết quả phân tích diễn ngôn tiếng Việt. Khái niệm 2.5 được xây dựng dựa trên

nghiên cứu của Verberne [89-92] và Delmonte [25]. Điểm mới của Khái niệm 2.5 là đưa thêm lập luận vào danh sách đối tượng có thể chứa câu trả lời cho câu hỏi "TẠI SAO".

CHƯƠNG 3. PHÂN TÍCH DIỄN NGÔN TIẾNG VIỆT Ở CẤP ĐỘ CÂU VÀ LIÊN CÂU THEO QUAN HỆ LÝ DO

3.1 GIỚI THIỆU BÀI TOÁN

Luận án đã nghiên cứu xây dựng ngữ liệu chú giải diễn ngôn cho văn bản ngắn tiếng Việt [CT.3] nhưng sự đồng thuận trong kết quả chú giải không cao. Vì thế, thay vì xác định các loại quan hệ diễn ngôn ở mức văn bản, bài toán được đặt ra cho luận án là xác định các quan hệ diễn ngôn thuộc nhóm quan hệ lý do theo **Khái niệm 2.4** ở cấp độ câu.

3.2 BÀI TOÁN PHÂN ĐOẠN EDU

Luận án xác định ranh giới của mỗi EDU theo định dạng bài toán phân đoạn EDU của hội nghị DISRPT [108]. Đơn vị diễn ngôn cơ bản (EDU) được xác định theo **Mục 2.1.1**. Việc xác định EDU sẽ được thực hiện ở cấp độ câu.

Để phân đoạn EDU tự động, luận án tinh chỉnh một mô hình pretrained BERT trên ngữ liệu chú giải EDU tiếng Việt. Ngữ liệu này được chuyển đổi từ ngữ liệu gán nhãn cú pháp cấu trúc ngữ đoạn NIIVTB [68] bằng cách xem xét các mẫu của các mệnh đề độc lập trong các cây cú pháp. Vấn đề này được trình bày trong **Phụ lục A.4** và đã được công bố ở [CT.6].

3.3 XÁC ĐỊNH QUAN HỆ DIỄN NGÔN THUỘC NHÓM QUAN HỆ LÝ DO Ở CẤP ĐỘ CÂU

Luận án nghiên cứu xây dựng mô hình phân lớp để nhận dạng quan hệ diễn ngôn nhưng kết quả không cao với F_1=0,4 [CT.3]. Nguyên nhân chủ yếu là ngữ

liệu chú giải cấu trúc diễn ngôn cho văn bản tiếng Việt còn quá ít vì cần người chú giải có kiến thức về phân tích diễn ngôn tiếng Việt. Bên cạnh đó, kết quả xây dựng tập dữ liệu chú giải câu trả lời cho câu hỏi "TẠI SAO" trong [CT.1] cho thấy rằng, mặc dù không thể hoàn toàn dựa vào các từ ngữ liên kết văn bản để xác định quan hệ diễn ngôn nhưng có thể sử dụng các từ ngữ này để nhận dạng các quan hệ diễn ngôn trong nhóm quan hệ lý do ở cấp độ câu với độ chính xác 78% [CT.1]. Điều này cũng phù hợp với việc dựa vào các từ ngữ liên kết để nhận dạng lập luận đời thường như đã nêu.

Vì lý do trên, luận án đã xây dựng văn phạm phi ngữ cảnh G_D=<Σ_D, N_D, P_D, S_D> bằng cách phân tích 500 câu có thể chứa quan hệ lý do được đánh dấu bằng từ ngữ liên kết. Ví dụ minh họa việc xây dựng văn phạm G_D, được trình bày ở **Phụ lục A.6** và kết quả tổng hợp tập luật sản sinh được trình bày ở **Phụ lục A.7**. Luận án dùng văn phạm G_D để phân tích quan hệ lý do ở cấp độ câu

3.4 XÁC ĐỊNH QUAN HỆ LÝ DO Ở MỨC LIÊN CÂU

Ở bài toán này, luận án dựa trên các từ chức năng ở đầu câu để xác định quan hệ lý do giữa các câu. Danh sách các từ chức năng ở mức liên câu được trình bày trong **Phụ lục A.5**

3.5 THỬ NGHIỆM VÀ ĐÁNH GIÁ

Phương pháp phân đoạn EDU dùng kiến trúc BERT của luận án có hiệu quả F_1=0,8 [CT.6] mặc dù chưa cao nhưng có thể sử dụng trong các thử nghiệm của luận án.

Kết quả phân tích quan hệ lý do cấp độ câu của luận án có độ đo F_1=0,7943 trên ngữ liệu thử nghiệm của luận án. Ngữ liệu này đã được phân đoạn EDU thủ công và tham số hóa thủ công các ngữ đoạn trong câu.

CHƯƠNG 4. PHƯƠNG PHÁP LẬP LUẬN TRÊN BIỂU DIỄN DẠNG VĂN BẢN TIẾNG VIỆT

4.1 PHƯƠNG PHÁP LẬP LUẬN TRÊN BIỂU DIỄN DẠNG VĂN BẢN THEO CƠ CHẾ LOẠI SUY

Luận án đề xuất phương pháp lập luận, được xây dựng trên cơ sở lập luận loại suy của Juthe [43], thực hiện trên biểu diễn dạng văn bản như sau:

Phương pháp lập luận đi từ các tiền đề P đến kết luận q trong phạm vi thông tin của một văn bản T và các tiền giả định [4] PS được thực hiện theo năm bước.

1) *Bước 1, phân tách văn bản thành tập E các EDU.*
2) *Bước 2, lần lượt với từng tổ hợp p_i các tiền đề trong P cho trước. Với mỗi p_i được xem là tiền đề:*
   - *Nếu $p_i \vDash q$ thỏa điều kiện là đối tượng loại suy của một lập luận $Arg \in PS$, qua Bước 4.*
   - *Ngược lại, chọn $c \in E, c \notin P$ sao cho $p_i \vDash c$ thỏa điều kiện là đối tượng loại suy của một lập luận $Arg \in PS$ theo lược đồ loại suy của Juthe[43]. Khi đó, $p_i \vDash c$ là một lập luận và cập nhật danh sách tiền đề $P = P \cup \{c\}$.*
3) *Bước 3, nếu không có lập luận mới được tạo ở Bước 2 thì đến Bước 5, ngược lại thì tiếp tục thực hiện Bước 2.*
4) *Bước 4, danh sách các lập luận được xác định trong Bước 2 biểu diễn một đồ thị. Kết quả của quá trình lập luận là một lập luận tương ứng với một đường đi từ các tiền đề ban đầu đến kết luận q đã cho. Dừng.*
5) *Bước 5, kết luận không thể lập luận để chứng tỏ kết luận q. Dừng.*

Phương pháp lập luận này được được xây dựng dựa trên phương pháp tính toán độ thuyết phục theo Khái niệm 2.3 theo kiến trúc mạng BERT dành cho bài toán NLI.

4.2 ỨNG DỤNG CỦA LẬP LUẬN TRÊN BIỂU DIỄN DẠNG VĂN BẢN

Luận án sử dụng phương pháp lập luận trên biểu diễn dạng văn bản để so khớp ngữ nghĩa của hai câu theo công thức (4.2) để tăng độ phủ khi áp dụng cho bài toán xác định cặp câu có nghĩa tương đồng. Bên cạnh đó, luận án cũng áp dụng phương pháp này để tính toán độ tương đồng ngữ nghĩa $Sim(p,q) \in [0,1]$ giữa hai câu p và q để so sánh và xếp hạng theo công thức (4.3). Công thức (4.3) tính toán trên độ thuyết phục của lập luận theo Khái niệm 2.3.

$$p \vDash q \vee q \vDash p \Rightarrow p = q \tag{4.2}$$

$$Sim(p,q) = \frac{Support(A_{pq}, PS) + Support(A_{qp}, PS)}{2} \tag{4.3}$$

Trong công thức (4.3):

- PS là tiền giả định, là cơ sở để tính độ thuyết phục theo Khái niệm 2.3.
- A_{pq} và A_{qp} lần lượt là lập luận có tiền đề p và kết luận q và lập luận có tiền đề q và kết luận p.

4.3 HUẤN LUYỆN MÔ HÌNH NHẬN DẠNG LẬP LUẬN TRÊN BIỂU DIỄN DẠNG VĂN BẢN VỚI KIẾN TRÚC BERT

4.3.1 Xây dựng bộ ngữ liệu huấn luyện

Luận án xây dựng bộ ngữ liệu *NLI* cho tiếng Việt, được đặt tên làm VnNewsNLI. Ngữ liệu VnNewsNLI được trình bày chi tiết ở [CT.4] và [CT.8]. Để tăng cường thêm ngữ liệu huấn luyện, luận án sử dụng thêm hai phiên bản dịch máy, gọi là VMNLI và VSNLI, từ tiếng Anh sang tiếng Việt của hai tập huấn luyện trong hai bộ ngữ liệu lần lượt là MultiNLI [99] và SNLI [13]. Việc dịch tự

động được thực hiện nhờ mô hình dịch máy Anh – Việt Helsinki[3] với điểm số BLEU đạt 37,2%.

4.3.2 Huấn luyện mô hình nhận dạng lập luận loại suy cho tiếng Việt

Mô hình lập luận loại suy cho tiếng Việt được huấn luyện theo kiến trúc mạng sử dụng BERT. Trong kiến trúc này, $\text{PhoBERT}_{\text{base}}$ [65] được sử dụng để sinh vector ngữ nghĩa của cặp tiền đề – kết luận.

4.4 ĐÁNH GIÁ MÔ HÌNH NHẬN DẠNG LẬP LUẬN LOẠI SUY TRÊN BIỂU DIỄN DẠNG VĂN BẢN

Mô hình lập luận loại suy M+News của luận án được đánh giá trên tập thử nghiệm của ngữ liệu VnNewsNLI đạt độ chính xác acc=0,9514. Mô hình M+News được chọn để sử dụng trong thử nghiệm đánh giá kết quả trả lời câu hỏi "TẠI SAO". Nhược điểm của mô hình này là chỉ áp dụng để tạo các lập luận đơn giản.

CHƯƠNG 5. MÔ HÌNH LẬP LUẬN ĐỂ TRẢ LỜI CÂU HỎI TẠI SAO

5.1 PHƯƠNG PHÁP LẬP LUẬN ĐỂ TRẢ LỜI CÂU HỎI "TẠI SAO"

Phương pháp lập luận để trả lời câu hỏi "TẠI SAO" của luận án như sau:

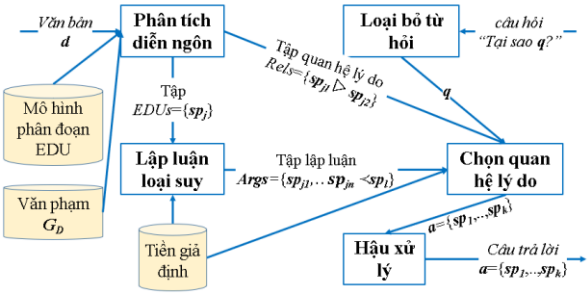
- Bước chuẩn bị: xây dựng tiền giả định [4] (presupposition) PRE gồm các cặp câu tương ứng với tiền đề và kết luận của một lập luận có hiệu lực.
- Bước 1: phân tích văn bản thành tập các EDU $P = \{p_i\}$ trong đó p_i là một EDU và cũng là một mệnh đề.
- Bước 2: phân tích diễn ngôn theo quan hệ lý do để xác định tập $Rels$ chứa các lập luận hoặc lời giải thích trong văn bản.

[3] https://huggingface.co/Helsinki-NLP/opus-mt-en-vi

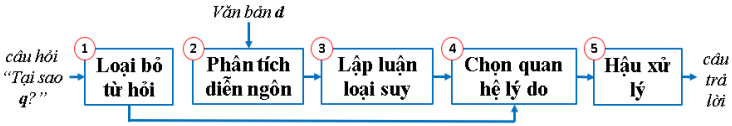
- Bước 3: xác định tập $Args$ chứa các lập luận được xác lập từ tập các mệnh đề P bằng cách áp dụng lược đồ lập luận loại suy của Juthe với tiền giả định PRE.
- Bước 4: Xây dựng đồ thị $TArg$ biểu diễn các lập luận và các lời giải thích trong tập $ArgRhe$ và $ArgAna$. Đồ thị $TArg$ có chiều của cạnh đi từ kết luận đến các tiền đề. Quá trình xác định lý do bằng cách duyệt qua đồ thị lý do chính là quá trình áp dụng quy tắc suy luận bắc cầu.

5.2 MÔ HÌNH LẬP LUẬN ĐỂ TRẢ LỜI CÂU HỎI "TẠI SAO"

Mô hình hệ thống, được thể hiện trong Hình 5.1, được đề xuất nhằm xử lý văn bản và câu hỏi "TẠI SAO" để trả về câu trả lời theo **Khái niệm 2.5**. Quy trình xử lý của hệ thống được thể hiện trong Hình 5.2.



Hình 5.1 Mô hình hệ thống lập luận để trả lời câu hỏi "Tại sao" dựa trên cách tiếp cận phân tích diễn ngôn tiếng Việt

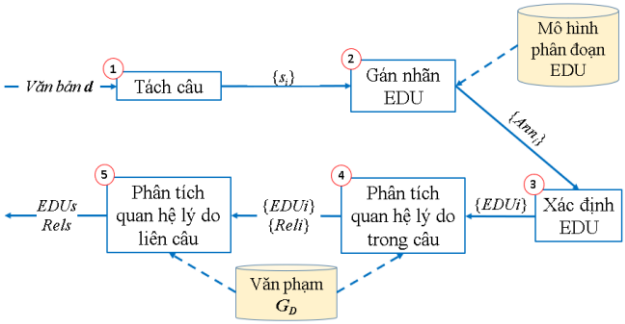


Hình 5.2 Quy trình xử lý câu hỏi của hệ thống lập luận để trả lời câu hỏi "Tại sao" dựa trên cách tiếp cận phân tích diễn ngôn tiếng Việt

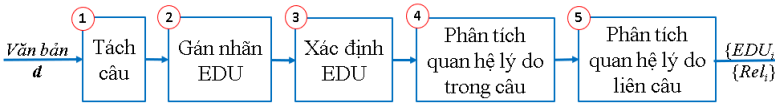
Đầu vào của mô hình là một văn bản d và một câu hỏi tại sao q. Đầu ra của mô hình là một câu trả lời $a = \{sp_1, sp_2, \dots, sp_k\}, sp_j \in d$.

5.2.1 Thành phần phân tích diễn ngôn

Phân tích diễn ngôn của văn bản theo quan hệ lý do được xây dựng trên kết quả nghiên cứu của luận án. Thành phần này có sơ đồ thiết kế theo Hình 5.3. Quy trình xử lý văn bản đầu vào d của thành phần này được thể hiện ở Hình 5.4.



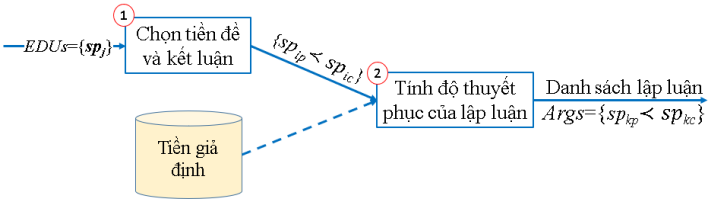
Hình 5.3 Sơ đồ thiết kế thành phần phân tích diễn ngôn theo quan hệ lý do



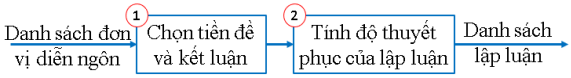
Hình 5.4 Quy trình xử lý văn bản của thành phần phân tích diễn ngôn theo quan hệ lý do

5.2.2 Thành phần lập luận loại suy

Tạo các lập luận gồm một tiền đề và một kết luận trong đó tiền đề và kết luận đều là một EDU của văn bản d. Thành phần này có sơ đồ thiết kế thể hiện ở Hình 5.6, có quy trình xử lý được trình bày trong Hình 5.7.



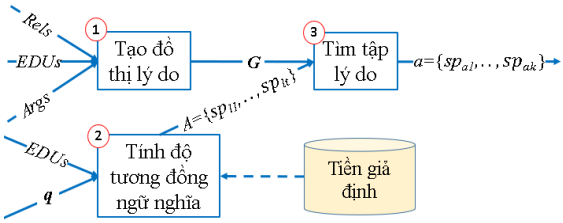
Hình 5.6 Sơ đồ thiết kế thành phần lập luận loại suy



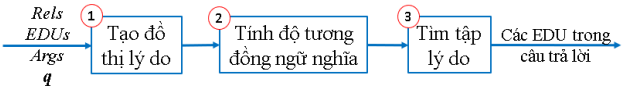
Hình 5.7 Quy trình xử lý của thành phần lập luận loại suy

5.2.3 Thành phần chọn quan hệ lý do

Chọn một tập $a = \{sp_{a1}, sp_{a2}, .., sp_{ak}\}$ chứa các EDU theo Khái niệm 2.5. Tập a thể hiện các ý có trong câu trả lời. Thành phần này có sơ đồ thiết kế thể hiện ở Hình 5.9, có quy trình xử lý được thể hiện ở Hình 5.10.



Hình 5.9 Thành phần chọn quan hệ lý do



Hình 5.10 Quy trình xử lý của thành phần chọn quan hệ lý do

5.2.4 Thành phần hậu xử lý

Rút gọn câu trả lời của mô hình bằng cách loại bỏ một số EDU là các tiền đề trung gian trong câu trả lời. Lý do là tập $a = \{sp_{a1}, sp_{a2}, \dots, sp_{ak}\}$ chứa các EDU trong câu trả lời nhưng đáp án thường chỉ gói gọn trong một EDU. Nếu dùng toàn bộ EDU từ thành phần Chọn quan hệ lý do sẽ làm điểm số F_1 của hệ thống giảm.

5.3 NGỮ LIỆU THỬ NGHIỆM

5.3.1 Ngữ liệu thử nghiệm để đánh giá mô hình

Các tập ngữ liệu thử nghiệm theo bài toán dạng rút gọn là tập VnYQA (100 mẫu) và ViYQuAD (316 mẫu). Mỗi câu hỏi trong tập VnYQA và ViYQuAD có câu trả lời là một chuỗi duy nhất.

Tập VnYNews được dùng trong thử nghiệm theo bài toán rút gọn nhưng câu trả lời có nhiều hơn một ý và các ý này không liên tục nhau trong văn bản. Thử nghiệm này nhằm thể hiện ưu điểm của phương pháp của luận án. Tập VnYNews gồm 203 mẫu thử nghiệm.

Các mẫu thử nghiệm trong tập VnYQA và ViYQuAD được chia thành ba nhóm gồm nhóm khó, nhóm trung bình và nhóm dễ dựa vào kết quả truy xuất câu theo ba mô hình gồm mô hình TF.IDF, mô hình tập hợp và mô hình NLI .

5.3.2 Ngữ liệu huấn luyện mô hình rút trích câu trả lời

Luận án dùng hai tập ngữ liệu để huấn luyện mô hình học sâu cho hệ thống end-to-end là vSQuAD và ViQuAD. vSQuAD là tập huấn luyện ngữ liệu SQuAD v1.1 [76] được dịch sang tiếng Việt bằng một chương trình dịch máy [CT.5]. Tập

ViQuAD là tập huấn luyện đã được loại bỏ những mẫu không có câu trả lời trong ngữ liệu UIT-ViQuAD v2.0.

5.4 CÁC CHƯƠNG TRÌNH ĐƯỢC THỬ NGHIỆM

Thử nghiệm được tiến hành với 5 chương trình IRYQA, QII-PhoBERT, UIT-PhoBERT, UIT-DistilBERT, UIT-XLMR, BERTYQA, OH-YQA và MHOPQA. Các chương trình thử nghiệm này sử dụng kiến trúc BERT thay cho các kiến trúc mạng học sâu khác để không có sự khác biệt nhiều giữa các chương trình về kỹ thuật tính toán do sự khác biệt của các kiến trúc mạng.

5.4.1 Chương trình IRYQA

Được cài đặt theo cách tiếp cận của Verberne [92]. Chương trình này dùng danh sách từ ngữ liên kết được xác định trong [CT.1] để nhận dạng các câu có cấu trúc nguyên nhân – kết quả; thành phần truy xuất câu sử dụng công thức tính độ tương đồng TF.IDF.

5.4.2 Chương trình QII-PhoBERT

Chương trình QII-PhoBERT, được cài đặt và thử nghiệm trong công trình [CT.5], sử dụng kiến trúc BERT của Devlin [26], là một hệ thống single-hop theo cách tiếp cận mạng nơron học sâu. QII-PhoBERT được tinh chỉnh từ mô hình pretrained $\text{PhoBERT}_{\text{base}}$ [65] trên ngữ liệu vSQuAD.

5.4.3 Chương trình UIT-PhoBERT

Chương trình này là phiên bản được huấn luyện trên ngữ liệu ViQuAD của QII-PhoBERT. UIT-PhoBERT được thử nghiệm nhằm so sánh kết quả trả lời câu hỏi "TẠI SAO" của kiết trúc BERT khi huấn luyện với ngữ liệu được được chú giải thủ công với kết quả của luận án.

5.4.4 Chương trình UIT-DistilBERT

Chương trình này là phiên bản của chương trình UIT-PhoBERT trong đó mô hình đa ngôn ngữ pretrained DistilBERT sử dụng thay cho PhoBERT

5.4.5 Chương trình UIT-XLMR

Chương trình này là phiên bản của chương trình UIT-PhoBERT trong đó mô hình đa ngôn ngữ pretrained XLM-R sử dụng thay cho PhoBERT.

5.4.6 Chương trình BERTYQA

Chương trình BERTYQA là chương trình thử nghiệm được cài đặt theo mô hình đề xuất của luận án nhằm đánh giá kết quả của phương pháp lập luận trong việc trả lời câu hỏi "TẠI SAO" của luận án. Chương trình được cài đặt theo mô hình đã đề xuất với các thành phần được cài đặt theo **Mục 5.2**.

5.4.7 Chương trình OH-YQA

Chương trình OH-YQA là chương trình $\text{OH-YQA}_{\text{sentence}}$ được cài đặt thử nghiệm trong [CT.7] theo hệ thống hỏi-đáp câu hỏi "TẠI SAO" của Oh và cộng sự [70]. Chương trình này được tinh chỉnh từ mô hình pretrained $\text{PhoBERT}_{\text{base}}$ [65] trên tập huấn luyện VNCE đã chú giải cấu trúc nguyên nhân – kết quả và đã được sử dụng ở [CT.7].

5.4.8 Chương trình MHOPQA

Chương trình này được cài đặt theo mô hình hệ thống FE2H của Li và cộng sự [52] vì hệ thống FE2H là một trong những hệ thống có kết quả tốt nhất khi đánh giá trên ngữ liệu HotpotQA [106]. Chương trình MHOPQA là một hệ thống multi-hop, được xây dựng theo cách tiếp cận nơron học sâu và được tinh chỉnh từ mô hình pretrained $\text{PhoBERT}_{\text{base}}$ [65].

5.5 THỬ NGHIỆM VÀ ĐÁNH GIÁ

5.5.1 Thử nghiệm bài toán trả lời câu hỏi "TẠI SAO" dạng rút gọn

5.5.1.1 Kết quả thử nghiệm

Thử nghiệm các chương trình IRYQA, QII-PhoBERT, UIT-PhoBERT, UIT-DistilBERT, UIT-XLMR, BERTYQA, OH-YQA và MHOPQA trên tập thử nghiệm VnYQA và ViYQuAD. Kết quả thử nghiệm được trình bày trong Bảng 5.4 và Bảng 5.5 cho thấy cách tiếp cận của luận án có hiệu quả trên mức trung bình nhưng không hiệu quả bằng cách tiếp cận single-hop tinh chỉnh trên mô hình pretrained $\text{PhoBERT}_{\text{base}}$ với ngữ liệu huấn luyện tiếng Việt.

Bảng 5.4 Kết quả thử nghiệm của các chương trình trên tập thử nghiệm VnYQA và ViYQuAD
(*giá trị lớn nhất và lớn thứ hai lần lượt được in đậm và in nghiêng*)

Chương trình	**VnYQA**		**ViYQuAD**	
	F_1(%)	**Tỉ lệ chứa đáp án (%)**	**F_1(%)**	**Tỉ lệ chứa đáp án (%)**
IRYQA	27,9	68,0	26,5	**37,0**
QII-PhoBERT	*52,3*	61,0	37,4	20,7
UIT-PhoBERT	**57,4**	*71,0*	**45,3**	*34,5*
UIT-DistilBERT	0,17	0,0	0,1	0,0
UIT-XLMR	17,9	3,0	10,2	5,6
OH-YQA	23,2	55,0	20,5	22,6
MHOPQA	47,5	60,0	*37,6*	14,1
BERTYQA	42,6	**80,0**	28,5	29,2

Bảng 5.5 Kết quả thử nghiệm của các chương trình theo nhóm câu hỏi trên tập thử nghiệm VnYQA và ViYQuAD
(*giá trị lớn nhất và lớn thứ hai lần lượt được in đậm và in nghiêng*)

Chương trình	**Nhóm khó**		**Nhóm trung bình**		**Nhóm dễ**	
	Số câu	**Tỷ lệ(%)**	**Số câu**	**Tỷ lệ(%)**	**Số câu**	**Tỷ lệ(%)**
Ngữ liệu thử nghiệm VnYQA						
IRYQA	0	0,0	9	34,6	**59**	**100,0**
QII-PhoBERT	0	0,0	*17*	*65,4*	44	74,6

UIT-PhoBERT	*3*	*20,0*	*17*	*65,4*	*51*	*86,4*
UIT-DistilBERT	0	0,0	0	0,0	0	0,0
UIT-XLMR	0	0,0	0	0,0	3	5,1
OH-YQA	1	6,7	11	42,3	43	72,9
MHOPQA	2	13,3	12	46,2	46	78,0
BERTYQA	**7**	**46,7**	**23**	**80,8**	50	84,7
Ngữ liệu thử nghiệm ViYQuAD						
IRYQA	1	0,6	**64**	**71,1**	**53**	**100,0**
QII-PhoBERT	18	10,4	25	27,8	23	43,4
UIT-PhoBERT	**33**	**19,1**	*41*	*45,6*	*36*	*67,9*
UIT-DistilBERT	0	0,0	0	0,0	0	0,0
UIT-XLMR	7	4,0	8	8,9	3	5,7
OH-YQA	22	12,7	25	27,8	25	47,2
MHOPQA	8	4,6	23	25,6	14	26,4
BERTYQA	*27*	*15,6*	34	37,8	32	63,4

5.5.1.2 Phân tích kết quả

Câu trả lời của chương trình BERTYQA thường dài hơn đáp án rất nhiều, dẫn đến độ đo F_1 thấp hơn các chương trình UIT-PhoBERT, QII-PhoBERT và MHOPQA.

Tỉ lệ chứa đáp án của chương trình BERTYQA cũng chỉ ở mức trên trung bình mặc dù tỉ lệ này khá cao ở tập thử nghiệm VnYQA. Tỉ lệ chứa đáp án phụ thuộc vào khả năng nhận dạng được quan hệ lý do giữa các EDU trong văn bản. Có hai điểm yếu của chương trình BERTYQA dẫn đến khả năng nhận dạng quan hệ lý do còn chưa tốt. Điểm yếu thứ nhất là phương pháp phân tích quan hệ diễn ngôn trong nhóm quan hệ lý do ở cấp độ câu chủ yếu dựa trên từ ngữ liên kết và kết quả phân tích chỉ đạt F_1= 0,7943 trên các mẫu có dùng từ ngữ liên kết. Điểm yếu thứ hai là lập luận loại suy không nhận dạng được lời giải thích mà không dùng từ ngữ liên kết. Điều này thể hiện rõ ở tỉ lệ chứa đáp án của chương trình BERTYQA (27,8%) và của chương trình OH-YQA (38,9%) trong nhóm câu hỏi khó ở tập thử nghiệm ViYQuAD.

Bên cạnh đó, kết quả của chương trình UIT-DistilBERT và UIT-XLMR rất thấp là do các mô hình pretrained DistilmBERT và XLM-R là những mô hình đa

ngôn ngữ, việc tính toán theo đặc trưng tiếng Việt không tốt bằng mô hình pretrained $PhoBERT_{base}$.

5.5.2 Thử nghiệm với điều kiện câu trả lời có nhiều hơn một ý

5.5.2.1 Kết quả thử nghiệm

Thử nghiệm các chương trình IRYQA, QII-PhoBERT, UIT-PhoBERT, BERTYQA, OH-YQA và MHOPQA trên tập thử nghiệm VnYNews. Kết quả thử nghiệm trình bày ở Bảng 5.6. Thử nghiệm này nhằm cho thấy điểm khác biệt của phương pháp lập luận để trả lời câu hỏi "TẠI SAO" dựa trên cách tiếp cận phân tích diễn ngôn tiếng Việt của luận án.

Bảng 5.6 Kết quả thử nghiệm của các chương trình với câu trả lời có nhiều ý trên tập thử nghiệm VnYNews
(*giá trị lớn nhất và lớn thứ hai lần lượt được in đậm và in nghiên*)

Mô hình	Số ý	Số câu	F_1(%)
IRYQA	34	32	14,3
QII-PhoBERT	46	46	18,9
UIT-PhoBERT	*61*	*57*	**24,1**
UIT-DistilBERT	3	3	0,1
UIT-XLMR	8	8	4,7
OH-YQA	34	30	15,5
MHOPQA	3	3	*17,9*
BERTYQA	**98**	**77**	17,2

5.5.2.1 Phân tích kết quả

Kết quả của chương trình BERTYQA trong Bảng 5.6 có thể xác định được 98 ý trong 77 câu trả lời, là kết quả cao nhất về số ý và số câu trả lời trong các chương trình. Nguyên nhân là chương trình BERTYQA xác định câu trả lời bằng cách xác định các nhóm gồm các lời giải thích và các lập luận có liên quan đến nhau và có ngữ nghĩa tương đồng với nội dung câu hỏi "TẠI SAO" nhất.

Kết quả trong Bảng 5.6 cho thấy điểm nổi bật của phương pháp lập luận để trả lời câu hỏi "TẠI SAO" dựa trên cách tiếp cận phân tích diễn ngôn của luận án

so với các cách tiếp cận khác là xác định được nhiều ý trong câu trả lời có nhiều hơn một ý.

5.5.3 Thử nghiệm vai trò của các thành phần trong mô hình

5.5.3.1 Kết quả thử nghiệm

Thử nghiệm các chương trình $BERTYQA_{Dsc}$, $BERTYQA_{Arg}$ và $BERTYQA_{Pos}$, lần lượt là phiên bản loại bỏ thành phần phân tích diễn ngôn, thành phần lập luận và thành phần hậu xử lý trong mô hình hệ thống của chương trình BERTYQA trên các tập thử nghiệm VnYQA, ViYQuAD và VnYNews. Kết quả các thử nghiệm này được trình bày trong các Bảng 5.7, Bảng 5.8 và Bảng 5.9.

Bảng 5.7 Kết quả thử nghiệm khi không sử dụng một trong các thành phần trong mô hình của luận án trên tập thử nghiệm VnYQA và ViYQuAD
(*giá trị lớn nhất và lớn thứ hai lần lượt được in đậm và in nghiên*)

Chương trình	**VnYQA**		**ViYQuAD**	
	F_1(%)	**Tỉ lệ chứa đáp án (%)**	**F_1(%)**	**Tỉ lệ chứa đáp án (%)**
$BERTYQA_{Dsc}$	18,0	24,0	20,9	18,8
$BERTYQA_{Arg}$	**43,8**	74,0	*27,0*	24,5
$BERTYQA_{Pos}$	11,0	**96,0**	20,8	**44,5**
BERTYQA	*42,6*	*80,0*	**28,5**	*29,2*

Bảng 5.8 Kết quả thử nghiệm khi không sử dụng một trong các thành phần trong mô hình của luận án theo nhóm câu hỏi trên tập thử nghiệm VnYQA và ViYQuAD
(*giá trị lớn nhất và lớn thứ hai lần lượt được in đậm và in nghiêng*)

Chương trình	**Nhóm khó**		**Nhóm trung bình**		**Nhóm dễ**	
	Số câu	**Tỷ lệ(%)**	**Số câu**	**Tỷ lệ(%)**	**Số câu**	**Tỷ lệ(%)**
Ngữ liệu thử nghiệm VnYQA						
$BERTYQA_{Dsc}$	2	13,3	9	34,6	13	22,0
$BERTYQA_{Arg}$	6	40,0	18	69,2	50	84,7
$BERTYQA_{Pos}$	**14**	**93,3**	**26**	**100,0**	**56**	**94,9**
BERTYQA	7	46,7	23	80,8	50	84,7
Ngữ liệu thử nghiệm ViYQuAD						
$BERTYQA_{Dsc}$	18	10,4	23	25,6	19	35,8
$BERTYQA_{Arg}$	12	6,9	*35*	*38,8*	31	54,5

BERTYQA$_{Pos}$	**56**	**32,4**	**50**	**55,6**	**36**	**67,9**
BERTYQA	*27*	*15,6*	34	37,8	*32*	*60,4*

Bảng 5.9 Kết quả thử nghiệm khi không sử dụng một trong các thành phần trong mô hình của luận án với câu trả lời có nhiều ý
(*giá trị lớn nhất và lớn thứ hai lần lượt được in đậm và in nghiêng*)

Mô hình	Số ý	Số câu	F_1
BERTYQA$_{Dsc}$	78	60	14,5
BERTYQA$_{Arg}$	46	38	*15,3*
BERTYQA$_{Pos}$	**257**	**147**	10,8
BERTYQA	*98*	*77*	**17,2**

5.5.3.2 Phân tích kết quả

Bảng 5.7 và Bảng 5.8 cho thấy thành phần phân tích diễn ngôn quan trọng hơn thành phần lập luận nhưng theo Bảng 5.9 thì thành phần lập luận loại suy quan trọng hơn thành phần phân tích diễn ngôn. Vấn đề này xảy ra vì thành phần phân tích diễn ngôn của luận án chưa thể phân tích diễn ngôn của toàn văn bản mà chỉ phân tích một số quan hệ diễn ngôn thuộc nhóm quan hệ lý do và bằng cách dựa trên từ ngữ liên kết. Vì thế, với những văn bản sử dụng nhiều từ ngữ liên kết thì vai trò của thành phần này quan trọng hơn.

Hiện tượng khi dùng thành phần hậu xử lý làm tăng điểm số F_1 nhưng làm giảm mạnh tỉ lệ chứa đáp án của chương trình vì đáp án là một số lý do trong các lý do đã xác định được nhưng luận án chưa có phương pháp hiệu quả để loại bỏ các lý do không quan trọng.

5.6 ƯU ĐIỂM VÀ NHƯỢC ĐIỂM CỦA MÔ HÌNH

Mô hình của luận án có ưu điểm là xây dựng được một đồ thị quan hệ lý do giữa các EDU trong văn bản từ kết quả phân tích diễn ngôn và lập luận loại suy. Mô hình xác định câu trả lời dựa trên đồ thị lý do này nên có thể xác định câu trả lời có nhiều ý mà không phải chỉ ra số bước tìm câu trả lời như các mô hình multi-hop.

Mô hình có nhược điểm là độ F_1 và tỉ lệ câu trả lời chứa đáp án thấp hơn mô hình dùng kiến trúc BERT khi thử nghiệm với các mẫu có câu trả lời chứa một ý duy nhất. Nhược điểm này xuất phát từ ba nguyên nhân. Nguyên nhân thứ nhất là việc dùng EDU làm đơn vị phân tích văn bản và cũng là một ý trong câu trả lời. Điều này làm cho độ đo F_1 của mô hình không cao khi đáp án là một ngữ đoạn ngắn nhất có thể. Bên cạnh đó, thành phần hậu xử lý cũng làm cho tỉ lệ câu trả lời chứa đáp án giảm khi chọn lựa ý để có câu trả lời ngắn gọn hơn. Nguyên nhân thứ hai là do phương pháp phân tích diễn ngôn của luận án chỉ áp dụng được cho một số quan hệ diễn ngôn ở mức câu và liên câu. Nguyên nhân thứ ba là do phương pháp lập luận của luận án sử dụng mô hình sử dụng kiến trúc BERT được huấn luyện trên ngữ liệu còn chưa nhiều và còn đơn giản về cấu trúc lập luận.

5.7 KẾT CHƯƠNG

Phương pháp lập luận để trả lời câu hỏi "TẠI SAO" dựa trên cách tiếp cận phân tích diễn ngôn tiếng Việt của luận án chỉ ở mức trên trung bình, không hiệu quả bằng các phương pháp dùng mạng nơron học sâu theo dạng single-hop cả về độ đo F_1 và tỉ lệ câu trả lời chứa đáp án.

Điểm nổi bật của phương pháp được nghiên cứu trong luận án là khả năng tìm ra các ý trong câu trả lời có nhiều ý. Các ý này được trình bày trong các chuỗi không liên tục trong văn bản. Bên cạnh đó, phương pháp của luận án có thể cho thấy quá trình xác định câu trả lời của mô hình hệ thống và giải thích được kết quả của từng bước. Đây cũng là điểm nổi bật so với việc sử dụng thuần túy các mô hình học sâu vì các mô hình học sâu như các hộp đen, chưa thể giải thích quá trình tính toán để xác định câu trả lời.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

KẾT LUẬN

Kết quả thực hiện của luận án cho thấy mục đích của luận án đã được hoàn thành với các đóng góp cũng là bốn kết quả chính của luận án.

Kết quả đạt được

Kết quả thứ nhất là phương pháp phân tích quan hệ diễn ngôn của văn bản tiếng Việt theo quan hệ lý do ở cấp độ câu và liên câu. Các công trình nghiên cứu liên quan đã công bố gồm [CT.1], [CT.3] và [CT.6].

Kết quả thứ hai là phương pháp lập luận dựa trên lược đồ lập luận loại suy của Juthe. Các công trình liên quan đã được công bố gồm [CT.4] và [CT.8].

Kết quả thứ ba là phương pháp lập luận để trả lời câu hỏi “TẠI SAO” dựa trên cách tiếp cận phân tích diễn ngôn tiếng Việt được nghiên cứu trên cơ sở hai phương pháp lập luận và phương pháp phân tích quan hệ diễn ngôn theo quan hệ lý do ở cấp độ câu và liên câu. Kết quả nghiên cứu này được công bố trong [CT.7] và các công trình liên quan là [CT.2] và [CT.5].

Kết quả thứ tư là mô hình lập luận để trả lời các câu hỏi "TẠI SAO" dựa trên cách tiếp cận phân tích diễn ngôn tiếng Việt. Kết quả nghiên cứu này được công bố trong [CT.7].

Ngoài các đóng góp chính, luận án chuyển đổi được một bộ ngữ liệu EDU-UNI chú giải đơn vị diễn ngôn cơ bản tiếng Việt từ ngữ liệu NIIVTB chú giải cú pháp tiếng Việt với 9.000 câu, một bộ ngữ liệu VnNewsNLI chú giải quan hệ giữa một cặp tiền đề và kết luận cho tiếng Việt với 32.000 mẫu, và ngữ liệu VnYNews dùng để thử nghiệm hỏi-đáp với câu hỏi “TẠI SAO” gồm 100 mẫu với câu trả lời có thể có nhiều hơn một ý.

HƯỚNG PHÁT TRIỂN

Hướng phát triển đầu tiên là xây dựng ngữ liệu chú giải cấu trúc diễn ngôn cho văn bản tiếng Việt và ngữ liệu chú giải cho bài toán suy luận với ngôn ngữ tự nhiên (NLI) tiếng Việt. Khi có lượng ngữ liệu chú giải cấu trúc diễn ngôn và ngữ liệu chú giải lập luận lớn và đa dạng, sự kết hợp phân tích diễn ngôn và lập luận có thể áp dụng để xây dựng ứng dụng đánh giá bài văn nghị luận tự động; bổ sung khả năng giải thích thắc mắc của khách hàng bên cạnh việc trả lời những câu hỏi về tên sản phẩm, công dụng, giá cả, v.v.; và tóm tắt văn bản bằng cách chọn những đơn vị diễn ngôn có vai trò hạt nhân và loại bỏ bớt những đơn vị diễn ngôn có thể suy ra từ những đơn vị diễn ngôn khác.

Hướng phát triển thứ hai của luận án là kết hợp logic hình thức và suy luận trên ngôn ngữ tự nhiên sử dụng mô hình học sâu để tăng độ chính xác cũng như khả năng áp dụng trong một miền rộng. Bởi vì có những quy tắc suy luận đơn giản, chẳng hạn A và ¬A có quan hệ mâu thuẫn, nhưng phải cần có rất nhiều mẫu để mạng nơron có thể ghi nhớ được quan hệ này.

Hướng phát triển thứ ba của luận án là mở rộng bài toán để tìm câu trả lời trong tập tài liệu lớn. Khi này, cần phải nghiên cứu thêm phương pháp đánh giá lập luận nào có giá trị hơn để được chọn và phương pháp tổng hợp câu trả lời. Hiện tại, câu trả lời được tìm trong một văn bản ngắn nên chưa có sự xung đột giữa các lập luận có liên quan đến một phát biểu.

Danh mục công trình nghiên cứu

[CT.1] Chinh Trong Nguyen, Dang Tuan Nguyen: Construction of Vietnamese Argument Annotated Dataset for Why-Question Answering Method. ICTCC 2016: 124-132. DOI: 10.1007/978-3-319-46909-6_12

[CT.2] Chinh Trong Nguyen, Dang Tuan Nguyen: Towards an Argument-Based Method for Answering Why-question in Vietnamese Language. NICS 2016: 130-134. DOI: 10.1109/NICS.2016.7725637

[CT.3] Chinh Trong Nguyen, Dang Tuan Nguyen: Towards Building Vietnamese Discourse Treebank. SoICT 2017: 63-69. DOI: 10.1145/3155133.3155200

[CT.4] Chinh Trong Nguyen, Dang Tuan Nguyen: Building a Vietnamese Dataset for Natural Language Inference Models. FDSE 2021: 185-199. DOI: 10.1007/978-981-16-8062-5_12

[CT.5] Chinh Trong Nguyen, Dang Tuan Nguyen: A Vietnamese Answer Extraction Model Based on PhoBERT. ACOMP 2021: 112-119. DOI: 10.1109/ACOMP53746.2021.00022

[CT.6] Chinh Trong Nguyen, Dang Tuan Nguyen: "Elementary Discourse Unit Segmentation for Vietnamese Texts". In Int. J. of Intelligent Information and Database Systems. 2022. DOI: 10.1504/IJIIDS.2022.124090 (**Scopus – Q4**)

[CT.7] Chinh Trong Nguyen, Dang Tuan Nguyen: "Building a Discourse-Argument Hybrid System for Answering Vietnamese Why-questions". In Computational Intelligence and Neuroscience. 2021. DOI:10.1155/2021/6550871 (**SCIE – Q1, IF: 3.633**)

[CT.8] Chinh Trong Nguyen, Dang Tuan Nguyen: "Building a Vietnamese Dataset for Natural Language Inference Models". In SN Computer Science. 2022. DOI: 10.1007/s42979-022-01267-x